Developmental Science

Developmental Science 16:6 (2013), pp 959–966 DOI: 10.1111/desc.12036

SHORT REPORT

Statistical word learning at scale: the baby's view is better

Daniel Yurovsky,[1] Linda B. Smith[2] and Chen Yu[2]

1. *Department of Psychology, Stanford University, USA*
2. *Department of Psychological and Brain Sciences and Program in Cognitive Science, Indiana University, USA*

Abstract

A key question in early word learning is how children cope with the uncertainty in natural naming events. One potential mechanism for uncertainty reduction is cross-situational word learning – tracking word/object co-occurrence statistics across naming events. But empirical and computational analyses of cross-situational learning have made strong assumptions about the nature of naming event ambiguity, assumptions that have been challenged by recent analyses of natural naming events. This paper shows that learning from ambiguous natural naming events depends on perspective. Natural naming events from parent–child interactions were recorded from both a third-person tripod-mounted camera and from a head-mounted camera that produced a 'child's-eye' view. Following the human simulation paradigm, adults were asked to learn artificial language labels by integrating across the most ambiguous of these naming events. Significant learning was found only from the child's perspective, pointing to the importance of considering statistical learning from an embodied perspective.

Research highlights

- Shows that statistical word learning scales.
- Demonstrates that the first-person view facilitates learning.
- Describes the ambiguity distribution of natural naming events.

Introduction

The infant's world is filled with objects with unknown names, names that must be learned by mapping auditory words onto objects in the visual scene. To do this, young learners must contend with significant uncertainty: names may be heard in the context of scenes containing multiple unknown objects. Understanding the nature of this uncertainty, and explaining how young learners nonetheless manage to learn object names, is a major theoretical problem in the study of early word learning (Markman, 1990; Tomasello & Barton, 1994; Smith & Yu, 2008).

One approach to this theoretical problem focuses on how learners reduce uncertainty *within* a single naming event. Although a label may be heard in the context of many objects, learners may not treat them all as equally likely referents. Instead, they may use social and pragmatic cues to rule out contenders to the named target (Baldwin, 1991; Bloom, 2000; Tomasello, 2003). Within this framework, it is quite plausible that infants might map a word to a referent only when ambiguity can be reduced to a single target object. Contexts with insufficient cues for the infant to rule out all contenders might not lead to an attempt at mapping. If this is correct, a significant proportion of the naming events young children experience may not contribute to learning (Tomasello & Farrar, 1986; Bloom, 2000).

An alternative approach assumes that the heavy lifting of uncertainty reduction is accomplished *across* instances. Because a label's correct referent likely co-occurs with it more consistently than do other objects, word–referent mapping could be accomplished by aggregating co-occurrence information across multiple individually ambiguous naming situations (Siskind, 1996; Yu & Smith, 2007). Cross-situational word learning has been demonstrated empirically in both adults (Yu & Smith, 2007; Smith, Smith & Blythe, 2011; Yurovsky, Yu & Smith, in press) and young children (Smith & Yu, 2008; Scott & Fischer, 2012). Further, computational analyses

Address for correspondence: Daniel Yurovsky, Department of Psychology, Stanford University, 450 Serra Mall, Stanford, CA 94305, USA; e-mail: yurovsky@stanford.edu

show that if uncertainty in the world is like uncertainty in laboratory experiments – e.g. referents can be individuated and identified across naming events – cross-situational word learning will scale in rate and size to human lexicons (Blythe, Smith & Smith, 2010; Vogt, 2012). This leaves an open question: what is the nature of real-world naming event ambiguity, and is it amenable to cross-situational learning?

One recent study found real-world naming events to be significantly more uncertain than those studied in laboratory experiments, and concluded that cross-situational learning from these experiences was unlikely. Medina, Snedeker, Trueswell, and Gleitman (2011) followed four young children around their homes and recorded natural parent-generated naming events. The audio in these events was replaced with artificial language labels, and adult participants were then asked to learn labels for common objects from the vignettes. Medina *et al.* (2011) found that the majority of vignettes were highly ambiguous, and that adults could not learn the labels by integrating information across these ambiguous events. Indeed, guesses about the referent for each label did not become more accurate over multiple naming events. If the kind of referential ambiguity experienced by young learners is like that captured in these videos of parent naming, cross-situational learning may not be a viable mechanism for real-world word learning.

However, a second set of studies suggests the opposite problem with our understanding of real-world naming event ambiguity: visual contexts for young learners may be significantly *less* ambiguous than previously hypothesized. Attaching a small camera to toddlers' foreheads, Smith and colleagues (Yoshida & Smith, 2008; Smith, Yu & Pereira, 2011; Yu & Smith, 2012) measured the first-person visual input received by toddlers during naturalistic parent–child interactions. Although multiple toys were available, and all were typically in view for parents, children's views were characterized by considerable information reduction – often focused on a single visually dominant object. Nonetheless, there was still uncertainty, though perhaps of a different kind: not all parent-generated labels referred to the dominant objects in these children's view (Yu & Smith, 2012).

Could the word–referent ambiguity in the child's first-person view be better suited to cross-situational word learning than the ambiguity in a third-person view of the same naming event? To address this question, we used Medina *et al.*'s method (developed by Gillette, Gleitman, Gleitman & Lederer, 1999), asking adults to learn word–referent mappings from natural child-directed naming events. However, in addition to recording parent–child interactions from a tripod-mounted camera, we recorded the same interactions from a camera on the child's forehead (Figures 1a, b). Learning across ambiguous naming events from the third-person perspective was then compared directly to learning from the same events from the child's first-person perspective.

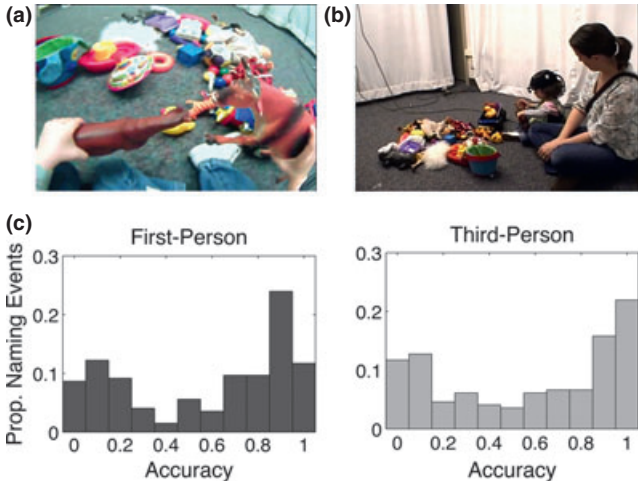

Figure 1 *Mother-child interactions were recorded from two views: a camera low on the child's forehead (a), and a tripod-mounted camera (b). Naming event accuracy was highly bimodal from both views (c). Only the most ambiguous events, as measured in Experiment 1, were used in Experiment 2.*

Experiment 1

Because the key theoretical idea behind cross-situational learning is that it enables learning via integration of information across individually ambiguous learning events, a critical first step is to determine the ambiguity of each naming event from the first- and third-person perspectives. Thus, we first recorded natural parent–child interactions and extracted naming events from the two views. The audio in each event was replaced with a beep, and adult participants were asked to guess the target of the mother's reference. Because each label was replaced by an identical beep, participants could not accumulate information about the likely referent across trials. Experiment 1 thus provides a measure of the ambiguity of each individual event, and also of the distribution of ambiguity across events from both views.

Method

Participants

The stimuli – child-directed naming events – were collected from play sessions in which four mothers

interacted with their four 2- to 2½-year-old children (mean age: 26;15, range: 25;12–27;10, two female). Each child received a small gift. For the experiment proper, 28 undergraduates participated in exchange for course credit. Half viewed naming events from the first-person perspective, and half viewed them from the third-person perspective.

Stimuli and design

Children and parents were asked to play naturally with toys while their interaction was recorded from a tripod-mounted camera and from a pinhole camera worn low on the child's forehead (see Appendix for details). After the head-camera and a vest carrying the power supply were put on the child, parent–child dyads played with the toys for approximately 10 minutes. Twenty-five toys were chosen to broadly sample the kinds of toys with which young children are likely to play – animals, cars and trucks, colored rings, a telephone, a baby doll, etc. (Figure 1a, b). Toys were arranged pseudorandomly in the center of the room when the play session began.

Each time a mother said the name of one of the toys, a vignette was created spanning from 3 seconds before the name to 2 seconds after. The audio was muted and a beep was inserted at the name's onset. If, in the natural interaction, the mother said the name again in the 2 seconds post-naming, another beep was inserted at this point and 2 more seconds of silent interaction were appended. The corpus consisted of 196 vignettes, 19 of which contained two beeps.

Procedure

Adult participants were informed that they would be watching videos of naming events from mother–child interactions. They were told that the beep in each video corresponded to a moment in the real interaction when the mother labeled one of the toys, and that they should guess the referent in each video. They were informed that multiple beeps in a single video always corresponded to a single referent. Adults then watched each vignette in the corpus once in random order, either from the first- or the third-person perspective (between subjects). At the end of each vignette, they were prompted to type in the most likely referent. Each adult first watched three vignettes from a pilot parent–child interaction to ensure that they understood the task. They were encouraged to guess on each vignette, and to describe objects as unambiguously as possible (e.g. 'white stuffed animal') if they could not determine their exact identity (e.g. bunny, sheep). This instruction was intended to minimize misidentification errors that could distort differences between the two camera views.

The instructions to participants differed from those in Medina *et al*. (2011), who did not specify that the target was an object nor did they accept guesses that did not exactly identify the target (e.g. 'purse' was considered incorrect for 'bag'). Pilot data from a group of participants not told that words referred to objects contained a significant proportion of guesses that were function words (e.g. 'the'), pronouns (e.g. 'it'), or onomatopoeias (e.g. 'whoosh'). We felt that these guesses were unlikely to be in the conceptual spaces of young children, and that our instructions increased the tenability of the human simulation hypothesis that adults were a proxy for young learners (Gillette *et al*., 1999).

Results and discussion

Since the focus of analysis is the population of naming events rather than the populations of adult guessers, we followed Medina *et al*. (2011) in aggregating guesses by vignette rather than by participant. Overall, the target object was identified almost 60% of the time, with similar accuracy from both views ($M_{1st} = .58, M_{3rd} = .58$, $t(195) = .26$, *ns*). However, ambiguity varied considerably across the 196 vignettes in the corpus.

As shown in Figure 1c, guess accuracy across vignettes was bimodal for both views. Approximately half of the naming events were highly ambiguous (19.1% were $\leq$ 10% accuracy) or highly unambiguous (29.6% were $\geq$ 90% accuracy). This distribution suggests that while many naming events may be unambiguous, a sizable proportion is likely to be opaque to single-instance learning mechanisms. These ambiguous naming events are exactly the kind of input over which cross-situational learning is hypothesized to operate. Can learners extract information from these ambiguous natural naming events? Experiment 2 addresses precisely this question, asking whether humans can learn word–referent pairings by integrating information across these ambiguous naming events, and whether learning depends on the perspective from which naming events are viewed.

Experiment 2

Experiment 2 asked participants to learn object names by aggregating evidence across the most ambiguous vignettes from Experiment 1. If cross-situational word learning does not scale to the ambiguity of natural naming events, then guess accuracy should not increase across vignettes (Medina *et al*., 2011). However, if the ambiguity characterizing first-person views is more

amenable to information aggregation than the ambiguity characterizing third-person views, then ambiguous naming events from the child's perspective should facilitate cross-situational learning even if learning fails from the third-person perspective.

Method

Participants

Forty-eight Indiana University undergraduates participated in exchange for course credit. Half of the participants watched vignettes from each perspective. None had participated in Experiment 1.

Stimuli and design

Stimuli for Experiment 2 were 20 naming event vignettes: four unique naming events for each of five different toys. For each vignette, the naming utterance was replaced with an artificial language label produced by a female native speaker of English. Six additional vignettes served as examples to acquaint participants with the task.

The selected vignettes were chosen such that the four naming events for each object came from at least two different parent–child dyads, and such that participant guessing accuracy for each event was $\leq$ 33% in Experiment 1. Experiment 1 guess accuracy for these vignettes was comparable across views (Figure 2). Six additional vignettes served as examples to explain the task, three of low ambiguity (Experiment 1 accuracy $\geq$ .8), and three of comparable ambiguity to those selected for cross-situational learning. Target referents for these example labels were different from the referents of cross-situational labels.

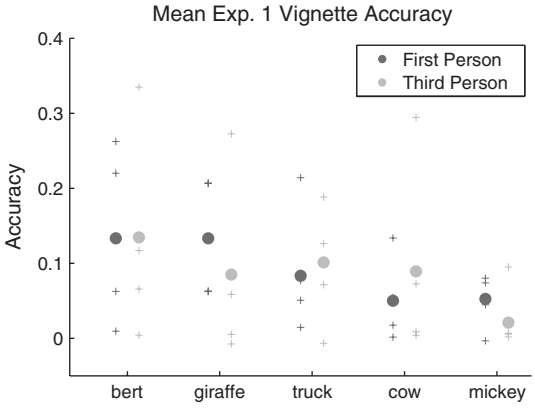

Figure 2 *Vignettes for Experiment 2 were chosen to be comparably difficult across views. Solid circles show mean Experiment 1 guess accuracy for the four vignettes used for each object in Experiment 2. Individual vignette accuracies are indicated by pluses, and are jittered with Normal(0, .001) noise for discriminability.*

Procedure

As in Experiment 1, participants were instructed to guess the object named on each trial. In addition, they were told that each unique artificial language label always referred to a consistent toy. Participants first watched three easy example vignettes for a single label. They then watched three difficult example vignettes for a different label. After participants demonstrated that they understood the task, they watched the 20 cross-situational vignettes in pseudorandom order such that the same label never occurred on successive trials. Twenty-four pseudorandom orders were created, and the order for each participant in the first-person view condition was yoked to the order for one participant in the third-person view condition.

Results and Discussion

Figure 3a shows guessing accuracy on each trial, averaged across the five individual labels. Accuracy on the first trial was low, and not significantly different across the two views ($M_{1st} = .12, M_{3rd} = .10$, $t(46) = .34$, *ns*). This validates the difficulty measure from Experiment 1, and verifies that vignettes in Experiment 2 were highly ambiguous. But while accuracies for the first trial were comparable across views, they diverged significantly with additional trials.

From the third-person view, accuracy did not increase significantly from the first vignette to any of the successive vignettes ($M_2 = .11$, $t(23) = .46$, *ns*; $M_3 = .16$, $t(23) = 1.43$, *ns*; $M_4 = .15$, $t(23) = .97$, *ns*). Further, guessing accuracy was uncorrelated with vignette number, indicating failure to learn across instances ($r = .12$, *ns*). Thus, the third-person view condition of Experiment 2 replicates Medina *et al.*'s (2011) results, showing no evidence of learning across ambiguous instances.

In contrast, accuracy in the first-person condition increased marginally from the first to the second vignette ($M_2 = .22$, $t(23) = 1.90$, $p = .07$), and was significantly higher on the third ($M_3 = .25$, $t(23) = 2.50$, $p < .05$) and fourth vignettes ($M_4 = .26$, $t(23) = 3.09$, $p < .05$). Further, vignette number and guess accuracy were significantly correlated ($r = .27$, $p < .01$).

Might differences in accuracy across views be due to differences in underlying learning mechanisms? Medina *et al.* (2011) distinguish between two qualitatively different mechanisms for cross-situational learning: *information accrual* vs. *single hypothesis* testing. In *information accrual* models, cross-situational learning

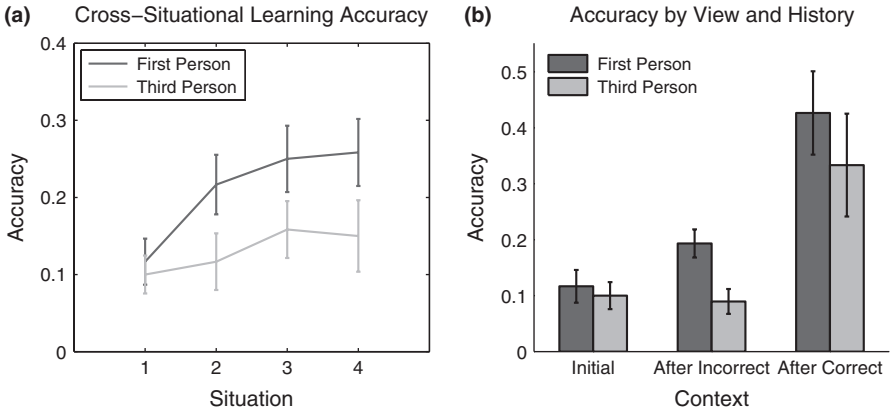

Figure 3 *(a) Naming event accuracy across instances from both views. Significant learning across instances was found only from the first-person view. (b) Naming event accuracy as a function of previous guess accuracy. In the third-person view, participants' guess accuracy improved only after correct guesses. However, in the first-person view, guess accuracy improved after incorrect guesses as well, suggesting accrual of information about multiple potential word–referent mappings.*

succeeds because learners track co-occurrence relationships between the words and objects in their input. Thus, from an ambiguous learning trial, an *information accrual* learner acquires information about the relationship between the word and multiple potential referents in the scene. In contrast, *single hypothesis* testers exposed to the same naming event remember only a single candidate object. On the subsequent naming event, this hypothesis is either confirmed and strengthened, or it is disconfirmed and the learner starts over as if from scratch. The *single hypothesis* model thus predicts that progress is made *only* after successful guesses; guess accuracy on a trial following an incorrect guess should be no higher than on the first learning trial. This prediction is upheld in Medina *et al*.'s (2011) data. We examine this prediction for learners from both views. If a participant guessed correctly on every trial, that participant was excluded from the analysis contingent on incorrect guesses. Similarly, if a participant guessed incorrectly on every trial, that participant was excluded from the analysis contingent on correct guesses.

Compared to accuracy on the first instance of a word, guesses made on an instance of that word following one on which a correct guess was made were more accurate from both the third- (M_{3rd} = .33, $t(41)$ = 2.76, p < .01) and first- (M_{1st} = .43, $t(44)$ = 4.0, p < .001) person views. Further, these accuracies were not significantly different from each other ($t(39)$ = .80, *ns*). Thus, learners in both views made progress after guessing correctly. After an incorrect guess, however, participants in the third-person view performed slightly, but not significantly, less well than after their first guess (M_{3rd} = .09, $t(46)$ = −.31, *ns*), but participants in the first-person view showed significant improvement (M_{1st} = .19, $t(46)$ = 1.98, p = .05). Further, accuracies from the two views were significantly different from each other ($t(46)$ = 3.08, p = .01). Thus, only from the first-person view did participants make significant progress after an incorrect guess, suggesting that the *single hypothesis* model is not a good account of learning from this view (Figure 3b).

Because the children in our study were 2 to 2½ years old, as in the original human simulation experiments (Gillette *et al*., 1999), they may have known the English-language labels for some toys from the free-play sessions and may also have used other linguistic information to navigate the visual scene. As the target of a child's gaze is a significant predictor of the target of mother's linguistic references (Frank, Tenenbaum & Fernald, in press), it is possible that the difference between views in our experiments is accounted for by different accessibility of the child's own knowledge. If children knew the English labels spoken in each vignette, they may have turned their heads in response, and these head-turns could have been easier to access from the first-person view. Since accuracy was comparable across views in Experiment 1, and for the first vignette for each label in Experiment 2, this is likely not the main driver of learning differences. Nonetheless, it could have contributed.

To test this possibility, post-referential head movement behavior was recorded by a naïve coder for each of the 20 cross-situational naming events in Experiment 2. The coder identified the target of the first attentional shift after the beep in each vignette, or alternatively indicated that no shift occurred. On 12 of the 20 naming events, children shifted their attention in the 2 seconds after the label was heard. However, only five of these attentional shifts were directed at the named object; the remaining seven were shifts to other toys in the room. Thus,

post-referential attentional shifts were not a good source of information in these vignettes.

To determine whether these shifts were nonetheless used differently across views, each vignette was assigned one of three values: no attentional shift (0), shift to correct object (1), or shift to incorrect object (−1). Average shift information for the four vignettes for each label was used to predict differences in final guess accuracy across views, but was found to be uncorrelated ($r = .01$, *ns*). Thus, children's own knowledge embodied in the videos cannot explain differences in learning across views.

We make two final notes about learning in Experiment 2. Although participants learned from the first-person perspective, (1) accuracy after the fourth vignette was still low, and (2) learning rate appears to decrease over exposures. These features may seem to suggest poor scalability for cross-situational learning. However, they are reliable features both of standard cross-situational learning (e.g. Kachergis, Yu & Shiffrin, 2012; Yurovsky *et al.*, in press), and of human (and animal) learning in general (e.g. Ebbinghaus, 1913; Rescorla & Wagner, 1972; Newell & Rosenbloom, 1981; for a review, see Heathcote, Brown & Mewhort, 2000). Although cross-situational word learning may have diminishing returns, word learners likely experience many more than four naming events.

To summarize, in the first-person, but not in the third-person view, guessing accuracy increased across multiple ambiguous instances, indicating integration of information about the referent of each label. Thus, viewing events from a first-person versus third-person perspective yields quantitative, and perhaps even qualitative, differences in cross-situational learning.

General discussion

Because children learn words so rapidly, acquiring more than 1300 words by 30 months of age (Mayor & Plunkett, 2011), many have argued that this learning cannot emerge from just the unambiguous naming events that children experience, but must also reflect the integration of information from less informative events (e.g. Siskind, 1996; Yu & Smith, 2007; Blythe *et al.*, 2010). Although the problem facing word learners in these experiments was simpler than the problem facing infant learners – they knew that labels referred to whole objects, they had to learn word–object rather than word–category mappings, they only had to learn five words, etc. (Medina *et al.*, 2011) – it was still orders of magnitude more complex than standard cross-situational word learning tasks (Yu & Smith, 2007; Smith *et al.*, 2011; Yurovsky *et al.*, in press). These results show that cross-situational learning can scale up to ambiguous real-world naming events. They also demonstrate for the first time that perspective matters: although participants saw identical, equally ambiguous naming events from both views, they successfully aggregated information only from the child's own view.

Why is the first-person view better? The clear implication is that all ambiguity is not created equal; instead the first-person view appears to contain usable regularities that are different from the third-person view. One source of these regularities may be the visual properties of the first-person view. Analyses of toddler's first-person views (e.g. Yu & Smith, 2012) and comparisons with third-person views (Yoshida & Smith, 2008) show differences in the dynamics of object foregrounding, differences in degree of clutter, and different patterns of visual salience. These differences may make contenders for the label's referent more memorable across trials, and they may limit the number of contenders – even when not clearly indicating the correct one. The accessibility of (potentially misleading) social cues may also differ between the views. For instance, mother's gaze does not reliably predict reference in many naming events (Frank *et al.*, in press). While analyses of head-camera views suggest that children access their mother's gaze infrequently (Franchak, Kretch, Soska & Adolph, 2011; Smith *et al.*, 2011), the third-person view makes gaze more readily available. Thus, information aggregation from the third-person view may be impeded by the participants' use of unreliable social cues. The results make clear that understanding the ambiguity of real-world naming events, and the information that learners exploit to aggregate information across such events, is critical to understanding the role of cross-situational learning in everyday word learning.

These results also provide new information about the distribution of ambiguity in natural naming events, showing it to be bimodal, with most naming events either unambiguous or highly ambiguous. Although Experiment 2 tested learning from only the most ambiguous events, the whole distribution is relevant to cross-situational word learning. Mounting evidence suggests that statistical speech segmentation, for instance, is bootstrapped by isolated words (Brent & Siskind, 2001; Lew-Williams, Pelucchi & Saffran, 2011). If word–referent learning operates similarly, and if information from ambiguous events is integrated with unambiguous events, then moments of referential clarity may play a critical role in modulating input to statistical word learning mechanisms (e.g. Yu, 2008; Frank, Goodman & Tenenbaum, 2009; Medina *et al.*, 2011).

Only regularities that make contact with children's sensory systems can affect their language learning. Consequently, the input to language learning must be understood from the learner's perspective. These experiments represent a critical first step to putting cross-situational learning on firmer ground by studying aggregation of information from the child's own view.

Acknowledgements

We are grateful to Amanda Favata who helped design the head-camera apparatus and collect mother–child interaction videos, and to Amber Matthew, Becca Baker, and Collin Caudell for help with human simulation experiments. We also thank all of the members of the Smith, Yu, and Shiffrin Labs as well as Jesse Snedeker, Dick Aslin, Hanako Yoshida, Anne Christophe, and two anonymous reviewers for their feedback on this project. This work was supported by a NSF GRF to DY and NIH R01HD056029 to CY and LBS.

References

Baldwin, D.A. (1991). Infants' contribution to the achievement of joint reference. *Child Development*, **62**, 875–890.

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

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

Brent, M.R., & Siskind, J.M. (2001). The role of exposure to isolated words in early vocabulary development. *Cognition*, **81**, B33–B44.

Ebbinghaus, H. (1913). *Memory: A contribution to experimental psychology*. New York: Teachers College Press.

Franchak, J.M., Kretch, K.S., Soska, K.C., & Adolph, K.E. (2011). Head-mounted eye tracking: a new method to describe infant looking. *Child Development*, **82**, 1738–1750.

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

Frank, M.C., Tenenbaum, J.B., & Fernald, A. (in press). Social and discourse contributions to the determination of reference in cross-situational word learning. *Language, Learning, and Development*.

Gillette, J., Gleitman, H., Gleitman, L., & Lederer, A. (1999). Human simulations of vocabulary learning. *Cognition*, **73**, 135–176.

Heathcote, A., Brown, S., & Mewhort, D.J.K. (2000). The power law repealed: the case for an exponential law of practice. *Psychonomic Bulletin & Review*, **7**, 185–207.

Kachergis, G., Yu, C., & Shiffrin, R.M. (2012). An associative model of adaptive inference for learning word–referent mappings. *Psychonomic Bulletin & Review*, **19**, 317–324.

Lew-Williams, C., Pelucchi, B., & Saffran, J.R. (2011). Isolated words enhance statistical language learning in infancy. *Developmental Science*, **14**, 1323–1329.

Markman, E.M. (1990). Constraints children place on word meanings. *Cognitive Science*, **14**, 57–77.

Mayor, J., & Plunkett, K. (2011). A statistical estimate of infant and toddler vocabulary size from CDI analysis. *Developmental Science*, **14**, 769–785.

Medina, T.N., Snedeker, J., Trueswell, J.C., & Gleitman, L.R. (2011). How words can and cannot be learned by observation. *Proceedings of the National Academy of Sciences, USA*, **108**, 9014–9019.

Newell, A., & Rosenbloom, P.S. (1981). Mechanisms of skill acquisition and the law of practice. In J. Anderson (Ed.), *Cognitive skills and their acquisition* (pp. 1–55). Hillsdale, NJ: Erlbaum.

Rescorla, R.A., & Wagner, A.R. (1972). A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. In A.H. Black & W.F. Prokasy (Eds.), *Classical conditioning ii: Current research and theory* (pp. 64–99). New York: Appleton Century Crofts.

Scott, R.M., & Fischer, C. (2012). 2.5-year-olds use cross-situational consistency to learn verbs under referential uncertainty. *Cognition*, **122**, 163–180.

Siskind, J.M. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, **61**, 39–91.

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

Smith, L.B., & Yu, C. (2008). Infants rapidly learn word–referent mappings via cross-situational statistics. *Cognition*, **106**, 1558–1568.

Smith, L.B., Yu, C., & Pereira, A.F. (2011b). Not your mother's view: the dynamics of toddler visual experience. *Developmental Science*, **14**, 9–17.

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Cambridge, MA: Harvard University Press.

Tomasello, M., & Barton, M.E. (1994). Learning words in nonostensive contexts. *Developmental Psychology*, **30**, 639–650.

Tomasello, M., & Farrar, M.J. (1986). Joint attention and early language. *Child Development*, **57**, 1454–1463.

Vogt, P. (2012). Exploring the robustness of cross-situational learning under Zipfian distributions. *Cognitive Science*, **36**, 726–739.

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

Yu, C. (2008). A statistical associative account of vocabulary growth in early word learning. *Language, Learning, and Development*, **4**, 32–62.

Yu, C., & Smith, L.B. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*, **18**, 414–420.

Yu, C., & Smith, L.B. (2012). Embodied attention and word learning by toddlers. *Cognition*, **125**, 244–262.
Yurovsky, D., Yu, C., & Smith, L.B. (in press). Competitive processes in cross-situational word learning. *Cognitive Science*.

Received: 28 March 2012
Accepted: 29 October 2012

Appendix

The head-camera's visual field was 90 degrees wide, providing a broad view of objects in the head-centered view at 10 frames per second. The camera's visual field does not capture the whole 170-degree toddler visual field, but is a good approximation (Smith *et al.*, 2011). The camera was attached to a headband that was tightened so that it did not move once set on the child. To calibrate the camera, the experimenter noted when the child focused on an object and adjusted the camera until the object was in the center of the image in the control monitor.

Because the head-camera moves with the child's head but not the child's eyes, its view of events may be momentarily misaligned with the direction of eye gaze. In a calibration study, Yoshida and Smith (2008) independently measured eye gaze direction (frame by frame via a camera fixated on the infant's eyes) and head direction and found that the two were highly correlated: 87% of head-camera frames coincided with independently coded directions of eye gaze. Moments of non-correspondence between head and eye directions in that study were generally brief (less than 500 msec). Thus, although head and eye movements can be decoupled, toddlers' tendency to align their head and eyes when interacting with objects suggests that the head-camera provides a reasonable measure of their first-person view.

The third-person camera was a Sony EVI-D70 camera mounted on a 3-foot-high tripod approximately 6 feet from the center of the toy room. The child and parent were free to move naturally around the room, but generally stayed between 3 and 10 feet away from the camera. The video was recorded at minimal zoom, providing a 48-degree viewing angle. If mothers or their children began to leave the frame, the experimenter panned the camera to follow their interaction. As in Medina *et al.* (2011), none of the vignettes tested included cases in which the toy was visible to the child but not to the camera.