Human Learners Integrate Visual and Linguistic Information in Cross-Situational Verb Learning

Yayun Zhang[1], Andrei Amatuni[1], Ellis Cain[2], Xizi Wang[3], David Crandall[2,3], Chen Yu[1,2,3]

yayunzhang@utexas.edu, andreiamatuni@utexas.edu, escain@iu.edu
xiziwang@iu.edu, djcran@iu.edu, chen.yu@austin.utexas.edu

[1]Department of Psychology, The University of Texas at Austin, USA
[2]Cognitive Science Program, Indiana University - Bloomington, USA
[3]Luddy School of Informatics, Computing, and Engineering, Indiana University - Bloomington, USA

Abstract

Learning verbs is challenging because it is difficult to infer the precise meaning of a verb when there are a multitude of relations that one can derive from a single event. To study this verb learning challenge, we used children's egocentric view collected from naturalistic toy-play interaction as learning materials and investigated how visual and linguistic information provided in individual naming moments as well as cross-situational information provided from multiple learning moments can help learners resolve this mapping problem using the Human Simulation Paradigm. Our results show that learners benefit from seeing children's egocentric views compared to third-person observations. In addition, linguistic information can help learners identify the correct verb meaning by eliminating possible meanings that do not belong to the linguistic category. Learners are also able to integrate visual and linguistic information both within and across learning situations to reduce the ambiguity in the space of possible verb meanings.

Keywords: verb learning, Human Simulation Paradigm, statistical learning, linguistic information, cross-situational learning

Introduction

Children's early productive vocabulary consist of significantly more nouns than verbs (Gentner, 1982; Goldin-Meadow et al., 1976). Abundant research has shown that verbs are harder to learn than nouns in general (Golinkoff et al., 1996; Halberda, 2003). Not only is it challenging for learners to discover the correct mapping between a verb and an action from the world, but it is also challenging to infer the underlying relational meaning that the verb encodes (Gentner, 1982; Snedeker, Gleitman, et al., 2004).

Noun learning is already hard, but verb learning is even harder. Imagine a child playing with some toys with her mother. The child picks up a toy phone and puts it near her ear. At this moment, she hears a new word from her mother. If the target word is a new noun, it is relatively easy for the child to infer that the correct referent for the heard noun is the toy phone. But if the target word is a verb, then there are many more candidate meanings that are embedded in the perceived event, such as "hold," "call," "answer," "put," and "talk." How do learners extract the precise meaning for the heard verb from many possible meanings?

Verb learning introduces a harder problem compared with noun learning. In the case of noun-object mappings, the referential uncertainty problem lies primarily in finding a target object from other potential objects. For verb-action mapping, on the other hand, identifying a target event is necessary but not sufficient to learn the meaning of a verb, which usually describes a relation within an event. Because an event can be conceptualized in terms of a multitude of relations, it is difficult to infer the meaning of a verb when there are many possible inductive generalizations that one can make from a single event (Childers et al., 2018; L. R. Gleitman and Gillette, 2017; Naigles, 1996).

One key mechanism that has been proposed to initially address this verb learning problem is the syntactic bootstrapping theory (L. Gleitman, 1990), which proposes that children use syntactic knowledge to decode verb meanings. For example, in English, verbs entailing one argument take intransitive frames (e.g. It fell), whereas verbs entailing two arguments take transitive frames (e.g. I dropped it). Children as young as 2.5 years old understand the syntactic structure of transitive and intransitive sentences and are able to assign different meanings to verbs in different syntactic structures (Arunachalam and Waxman, 2010; Fisher, 1996; Naigles, 1990, 1996; Yuan et al., 2012). Even though syntactic bootstrapping could help children narrow down a verb's meaning by providing linguistic constraints, these initial links children make are still not enough information to lead to the precise meaning. This is because it is very likely that among all possible meanings presented in the naming moment, more than one meaning can be described using verbs that fit the same syntactic structure. Therefore, instead of focusing on how children's developing verb knowledge is solely explained by their syntactic knowledge extracted from a single event, researchers have also started to ask whether children use cross-situational information to learn word meanings from multiple events.

Many studies have shown that both children and adults are good at using cross-situational statistics to find word-referent mappings (Horst et al., 2010; K. Smith et al., 2011; L. Smith and Yu, 2008; Trueswell et al., 2013; Yu and Smith, 2007). The basic idea behind cross-situational learning (CSL) is that when language learners are presented with multiple referents and multiple words in one naming moment, they are unable to decide which word maps onto which object. However, if learners keep track of multiple mappings where the same ob-

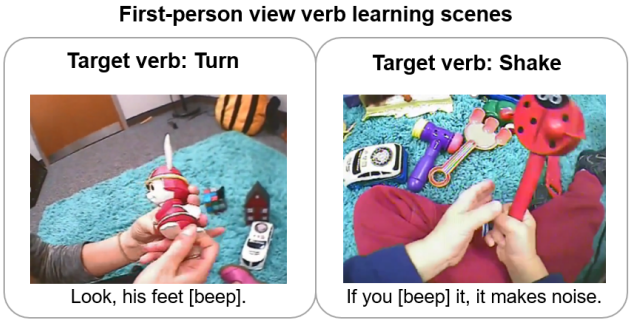

Figure 1: Screenshots from naturalistic toy-play showing the child's egocentric view at the moment when mothers named the verb "turn" and "shake." In conditions where linguistic information is provided, participants also hear the naming utterance with the target verb replaced by a beep.

ject co-occurs with other objects in other learning moments, the correct word-referent mappings will eventually emerge. Smith and Yu (2008) found that 12 to 14-months-olds could successfully associate six object names with their corresponding objects in a cross-situational learning task with thirty training trials. Researchers have also demonstrated cross-situational learning in verbs (e.g., Childers and Paik, 2009; Scott and Fisher, 2012; Waxman and Gelman, 2009). For example, Childers and Paik (2009) found that 2- to 3-year-olds were able to learn novel verbs by observing multiple visual events with different objects preserving the same action. However, most verb cross-situational learning studies used simple body movements as the target actions. These actions are concrete whole events that are very different from learners' experience of real-world events. In a naturalistic learning environment, it is the learners' job to "package" the elements of meanings to map onto verbs.

Determining the meaning of a verb by observing real world events is profoundly difficult. In Gillette et al.'s classic "human simulation" study, adult participants were asked to watch video clips of mothers interacting with their children. Each video clip contains moments when mothers uttered either a noun or a verb. The sound of each video was muted, and a beep was inserted at the onset of the target word. Participants were asked to guess which word the mother had said, indicated by the beep after each video. Although participants presumably had perfect conceptual knowledge about the information presented, they were only able to guess 27% of the nouns and 8% of the verbs correctly. Similar verb disadvantage has been observed using HSP with 7-year-olds (Piccin and Waxman, 2007). These findings suggest that real-world verb learning situations are inherently ambiguous and many verbs can be used to describe the same situation. Thus, inferring the precise meaning of a heard verb can be very challenging (Gillette et al., 1999).

Although multiple mechanisms have been proposed to explain how children learn verbs from the real world, it is still not clear whether learners are able to integrate visually grounded information from multiple naturalistic learning scenes to gradually identify the correct verb meaning and whether linguistic information can provide additional constraints to facilitate this verb learning process. To test these ideas in the current study, three experiments were conducted using the Human Simulation Paradigm (HSP). In Experiment 1 (baseline), we extracted verb naming instances from parent-child joint play (Figure 1) and quantified the degree of ambiguity in those instances by asking participants to guess the verb being uttered in each instance. The results from Experiment 1 were used as baseline measures for subsequent experiments. Experiment 2 (linguistic) aimed to understand whether linguistic information played a role in reducing in-the-moment referential ambiguity. The same set of videos used in Experiment 1 were used. Instead of using muted videos, we presented learners with the entire sentences that parents used in those naming moments except the target verbs, providing learners additional linguistic information. Experiment 3 focused on verb learning from multiple learning instances. Experiment 3a (CSL) examined whether learners could extract the correct verb meaning from multiple learning instances using visually grounded information only. The same videos from Experiment 1 were chunked into blocks with all instances in a given block referring to the same target. Participants were asked to watch each video and provide their best guess based on both visual information extracted from the current trial and statistical information accumulated from previous trials. Experiment 3b (CSL + linguistic) used the same set of blocked videos as in Experiment 3a, but with additional linguistic information. The goal of Experiment 3b was to examine whether learners could integrate visual and linguistic information together cross-situationally to identifying verb meanings.

We used the child's egocentric view extracted from naturalistic parent-child toy play contexts to closely simulate young learners' learning environment. Children's views may contain unique visual properties that provide different information in guiding referent selection compared to third-person or adults' views (Yurovsky et al., 2013).

Experiment 1

Experiment 1 was designed to explore whether learners were able to extract verb meanings from visually grounded information only.

Method

Participant. Forty-six undergraduate students recruited through the university subject pool (30 females, M = 19.22 years old, SD = 0.94) were included in the final analyses.

Stimuli. The video corpus included thirty-two parent-child (child age: M = 19.07 months old, SD = 3.14, range: 12.3-25.3 m.o.) dyads' play sessions, in which the dyads were instructed to play for ten minutes with a set of toys as they naturally would at home. The play session was recorded from the child's perspective using a head-mounted camera (Figure

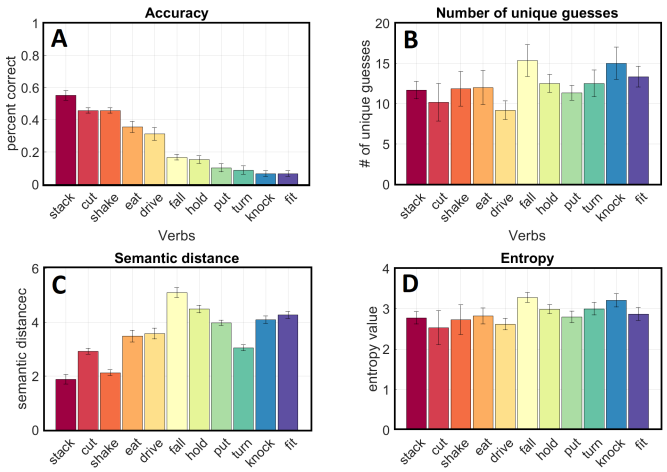

Figure 2: Experiment 1 (baseline): bar graphs showing results for four different measures.

1). From these play sessions, we transcribed parent speech and then used the transcriptions to identify the verb naming moments during the interactions. Among all of the verbs transcribed, we focused on concrete action verbs with visually grounded verb meanings.

The target referents consisted of 11 action verbs ("eat," "stack," "knock," "fit," "drive," "cut," "fall," "turn," "put," "hold," "shake") that commonly occur in toy-play interactions. Sixty-six videos were selected centered around a naming-moment. All videos were 5 seconds long, with the naming onset occurring at exactly three seconds. The original sound for each video was muted and a beep was played at the onset of the target verb to obscure the labelling event. Four additional videos with varying difficulties were included as training examples before the start of the experiment to ensure that participants understood the task.

Instructions and Procedure. The experiment took 20 minutes to complete. Participants were instructed to carefully watch short videos of parents playing with their children and to provide a guess for the intended verb at the moment of parent naming as indicated by the beep. They were told that the target verbs were all concrete action verbs, and therefore to enter correctly-spelled English concrete action verbs in the present tense. Each video was only played once, and participants had 20 seconds to enter a guess after watching each video. Feedback was not provided.

Results

Participants' individual guesses were corrected for incorrect spelling and tense. For data analyses, we derived four different measures to examine learning performance: 1) accuracy: whether learners could identify the correct target verb; 2) unique number of guesses: how many different verbs learners guessed for a given trial; 3) semantic distance: how close their guesses were to the correct target verb; 4) entropy: how learners' guesses were distributed.

Accuracy. Learners successfully guessed the target verb 22.6% of the time (*SD* = 0.08). This number is higher than the 8% verb learning accuracy reported in Gillette, Gleitman, Gleitman and Lederer (1999). This difference could be due to the different types of stimuli used (first-person vs. third-person view videos). Egocentric views may provide information with unique visual properties that can help learners constrain the meaning space. In addition, we found accuracy differences across verbs (Figure 2A). For example, verbs like "stack," "cut," and "shake" all had over 40% accuracy, whereas verbs like "put," "hold," and "fit" all had accuracy below 20%. This pattern is in line with previous work suggesting that the concepts verbs encode fall along a continuum of abstractness (Maguire et al., 2006). Although we only selected 11 relatively concrete action verbs that are commonly used in toy-play interactions, there were still differences in terms of the abstractness among these verbs. Verbs that represent concrete actions with more well-defined "shape" tend to be easier to learn.

Unique number of guesses. Despite this slight increase in learning accuracy, guessing the correct verb solely from visual information was still quite challenging. To further understand what learners guessed, we compiled all participants' guesses on a single trial and measured the number of unique guesses learners chose. As shown in Figure 2B, we found that on average, learners guessed about 12 different verbs (*SD* = 4.19) per trial. Consistent with what we hypothesized earlier, learners tended to pick different verbs to present the extracted meanings, suggesting a large search space for the correct meaning. Given this large search space, were these guesses semantically related to the correct target verb? It could be the case that learners were able to locate the right semantic space, but instead of using the exact target verb uttered by the parent to represent the meaning, they could choose another similar verb. If this is the case, then the guesses learners made should be closely related to the target verb.

Semantic distance. We then measured the semantic distances between target verb and each of the participants' guesses. We recruited another eighty participants from Amazon Mechanical Turk (26 females, *M* = 39.93 y.o., *SD* = 10.74). They were asked to rate 494 non-exhaustive verb pairs gathered from participants' guesses from the HSP task. Specifically, participants were told to rate two verbs as being similar if they can both be used to describe a concrete action in the same context. We found that on a scale of 1 to 7 (1 = very similar, 7 = very dissimilar), participants' guesses fell on the dissimilar end (*M* = 5.17, *SD* = 1.41), meaning that it is not the case that learners are using different verbs to describe the same meaning; instead, they may extract very different meanings from the same scene. It is still challenging for learners to locate the relevant semantic space using visual information alone (Figure 2C).

Entropy Lastly, we used entropy to quantify the distribution of participants' guesses (Figure 2D). Entropy is a mea-

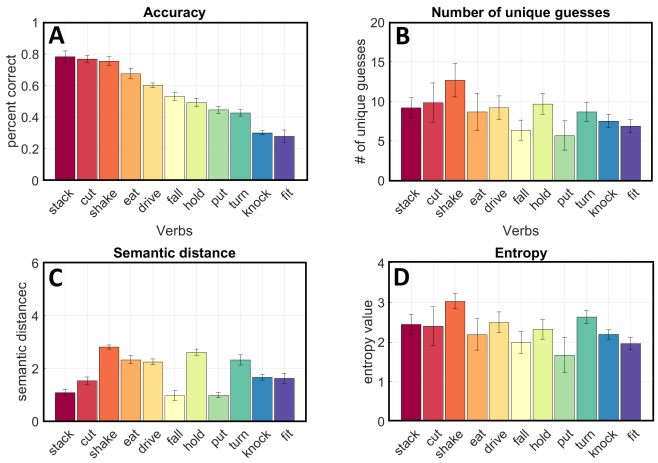

Figure 3: Experiment 2 (linguistic): bar graphs showing results for four different measures.

sure of uncertainty given a distribution,

$$H(x) = -\sum_{x=1}^{n} P(x)\log_2 P(x),$$

where *P(x)* represents the proportion of time verb x is guessed. This measure takes into account both the total number of verbs guessed and how often they were guessed. Lower entropy value means higher certainty. For example, for a given trial, if half of the learners chose verb A and half chose verb B, entropy is 1.0. If 75% of learners chose verb A and 25% chose verb B, then entropy drops to 0.81, indicating greater certainty. If learners chose 3 guesses equally often, then entropy increases to 1.56, suggesting lower certainty compared to an even distribution between two verbs. In Experiment 1, entropy was relatively high (M = 2.87, SD = 0.54), meaning that different participants tended to extract different verb meanings from each learning trial.

Results from Experiment 1 demonstrated that visual information in egocentric views was helpful for learning. More concrete verbs tend to be easier to learn than less concrete ones. However, the amount of guidance provided by visual information alone is quite limited as the search space for the target meaning remains quite large. Experiment 1 results also serve as a baseline for future comparisons.

Experiment 2

We next investigated how linguistic information could help reduce referential uncertainty in individual trials.

Method

Participants. Sixty-one participants recruited through Amazon Mechanical Turk (27 females, M = 40.17 y.o., SD = 9.83) were included in the final analyses.

Stimuli. The same videos used in Experiment 1 were used. To create linguistic information, we transcribed the original speech for each video and recreated them using Amazon Polly's text-to-speech natural speech synthesizer. For each trial, we only included the complete sentence containing the target verb (i.e., "Can you [beep] it?"; "[beep] the baby to bed.").

Instructions and Procedure. The procedure for Experiment 2 is similar to that of Experiment 1, with the addition of the linguistic information during the trial.

Results

We conducted similar analyses using the same four measures described in Experiment 1. We used the lme4 package in R to perform a series of mixed-effects analyses comparing each of the four measures between conditions, with random effects of subjects and items: accuracy $\sim$ baseline + (1 | subject) + (1 | item).

Accuracy. When presented with both visual and linguistic information, participants were more likely to guess the correct target verb (M = 0.54, SD = 0.09) compared to visual information only (β = 2.07, $p < 0.001$). Learning performance improved for all the verbs tested. Among the 11 verbs, "fall" had the largest amount of improvement of 36%, and "fit" had the smallest amount of improvement of 21%.

Unique number of guesses. Linguistic information also helped learners significantly narrow down the semantic search space by eliminating verbs that did not fit in the linguistic frame. The number of unique guesses dropped from around 12 (SD = 4.20) in Experiment 1 to around 8 (SD = 4.15) in Experiment 2 (β = -3.70, $p < 0.001$).

Semantic distance. Similarly, as the search space was becoming smaller and more constrained, we also found that verbs that learners guessed were also semantically closer to the target verb (β = -1.62, $p < 0.001$).

Entropy. The entropy value also dropped significantly to 2.30 (SD = 0.76), indicating higher certainty (β = -5.71, $p <$ 0.001).

Results from Experiment 2 suggest that linguistic information facilitates learning by helping learners narrow down the space of possible meanings in individual trials.

Experiment 3

Extending beyond individual trials, Experiment 3 examined whether learners can integrate meanings extracted from multiple scenes to gradually converge on the correct one.

Method

Participants. Twenty-eight undergraduate students recruited through the university subject pool (14 females, M = 20.03 y.o., SD = 3.08) participated in Experiment 3a. Thirty participants (13 females, M = 36.79 y.o., SD = 8.31) recruited through Amazon Mechanical Turk participated in Experiment 3b. Both groups used the same program to complete the task.

Stimuli. To provide additional statistical information to learners, we grouped the same 66 videos used in previous

experiments into 11 blocks with 6 trials in each block all referring to the same target verb. This design allows learners to utilize information across multiple trials to identify one verb meaning. Exp. 3a used videos containing visual information only and Exp. 3b used videos containing both visual and linguistic information.

Instructions and Procedure. Participants were informed that they would be watching blocks of videos where all of the videos within a block were naming the same target verb. Throughout the trials, they could change their guess within a block at any given trial, and were also allowed to enter the same answer if they believed their previous guess was correct. After each block, a prompt notified participants that the next trials belonged to a new block. In all experiments, participants were not allowed to go back to change previous answers and were not given any feedback during the experiment.

Results

To examine the role of statistical information, we present our results as three comparisons. To formally test the improvement over trials using lme4 package in R, we fit a mixed-effects logistic regression predicting each of the four different measures from trial number and baseline accuracy from Experiment 1 while taking into account the random intercepts for each subject and each item (mixed effects model: accuracy $\sim$ trial + baseline + (1 | subject) + (1 | item)).

Comparison 1: baseline vs. CSL. This comparison focuses on how statistical information helps learning when trials only contain visual information. As shown in red in Figure 4, we found a significant main effect of trial number (β = 0.25, $p = 0.01$) and baseline accuracy (β = 4.99, $p < 0.001$), suggesting significant learning across trials. One interesting pattern we found was that although learners provided guesses that are semantically closer to the target in later trials (β = 0.32, $p < 0.01$), they did not choose fewer unique guesses (β = -0.12, *ns*). This finding suggested that although learners were gradually shifting towards a more relevant search space, there was still a high degree of uncertainty in terms of what words learners use to represent the extracted meaning. It is likely that in later trials, learners have accumulated enough information to converge on a meaning but instead of using the exact target verb to represent that meaning, they instead chose a word that is semantically close to the target verb. When there were no other information available to further refine that space, different participants may choose different verbs and stay with their choices.

Comparison 2: Linguistic vs. Linguistic + CSL. When linguistic information was added to both the baseline and the CSL conditions, we also found learning improvement across trials (main effect of trial order: β = 0.38, $p < 0.001$, main effect of baseline accuracy: β = 2.98, $p < 0.001$). Similar to comparison 1, participants' guesses became closer to the correct meaning (semantic distance: $\beta = 0.29$, $p < 0.001$), but there was still variability in the words learners chose (number of unique guesses: β = 0.83, *ns*). Although each individual trial already contained rich visual and linguistic information, learners still benefit from seeing multiple events containing the same meaning in a sequence (blue lines in Fig 4).

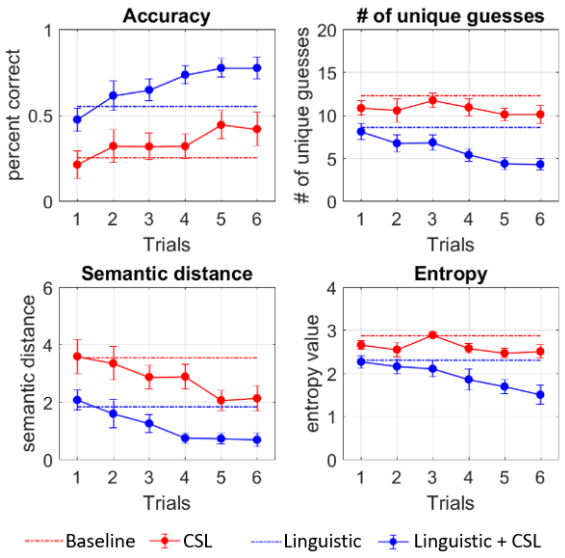

Figure 4: Line graphs showing how learners' guesses change across trials using four different measures.

Comparison 3: CSL vs. Linguistic + CSL. Lastly, we compared the two experiments containing CSL information (red vs. blue solid lines in Figure 4). We know from results in Experiment 2 that linguistic information had an additive effect on learning as it helps learners narrow down their search space. Here we found further evidence confirming that linguistic information improved learning significantly (accuracy: β = 0.93, $p < 0.001$; Number of unique guesses: β = 2.37, $p < 0.001$; semantic distance: β = 0.80, $p < 0.001$) across trials within a block. As shown in a concrete example of the target verb "eat" in Figure 5, we plotted histograms of participants' guesses at three time points for both Exp.3a and 3b. We see that within each experiment, early trials tend to have many more unique guesses than later trials, and even when learners did not eventually guess the correct verb "eat," they still identified semantically related words such as "bite" or "lick." Across the two experiments, we see a clear benefit of linguistic information regardless of trial positions, suggesting that linguistic and statistical information work together in constraining the meaning space.

In Experiment 3, we found strong evidence that when learners were provided with statistical information from a series of naming moments, they were more likely to find the correct verb meaning through information aggregation compared to observing individual moments alone.

Discussion

Our results suggested that visually grounded information perceived from the child's egocentric view was helpful for verb

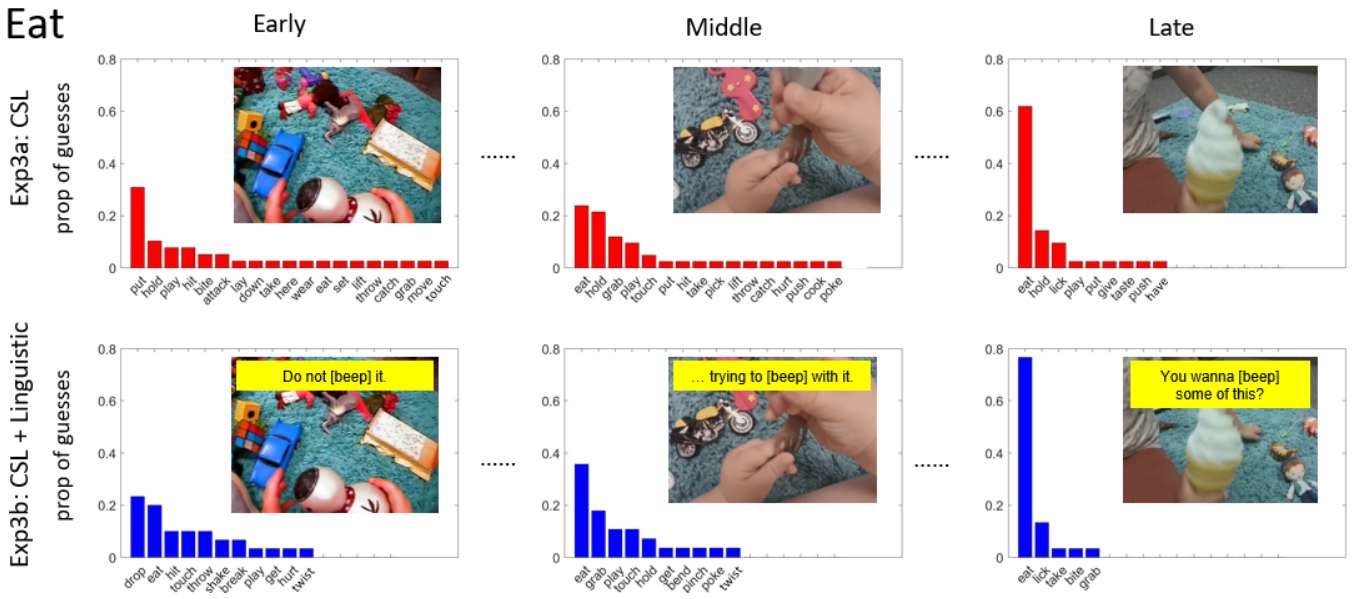

Figure 5: Histograms of participants' guesses at three time points during Exp. 3a and Exp. 3b.

learning. However, visual information alone is insufficient for learners to discover the precise meaning of a verb because a perceived action usually affords multiple meanings that are semantically close to one another. Linguistic information in the moment provides an additional source in the process of seeking the meaning of a verb. Moreover, beyond information embedded in individual learning trials, statistical information across trials could also facilitate verb learning. We show that learning performance was the best when both linguistic and statistical information were provided and used, suggesting that linguistic information within individual learning situations can be seamlessly integrated with cross-situational learning across multiple learning situations.

Our finding is in line with the classic syntactic bootstrapping theory suggesting that syntactic information provides learners with initial constraints about sentence structure and subsequently helps them reduce in-the-moment ambiguity (Gillette et al., 1999). The syntactic bootstrapping theory has been used to explain fast mapping through a single encounter of a word, suggesting that linguistic information can be critical for children to readily map words to referents in the world with only minimal exposure (e.g., Carey and Bartlett, 1978). Although syntactic bootstrapping has focused on the role of linguistic information in finding word-referent mappings, linguistic information is not the sole input used in word learning and it is questionable whether linguistic information alone is sufficient to "solve" the learning problem in a single learning moment. Because even after learners with sufficient syntactic knowledge successfully identify the target action event, they may still have difficulty proposing an effective hypothesis regarding verb meaning due to the "packaging" problem inherent in verb learning. Beyond individual learning moments, cross-situational information can always reduce the search space in the process of seeking the meaning of a word. The reduction on candidate meanings in individual moments narrows down statistical information which in turn allows learners to better integrate information across situations. By doing so, in-the-moment linguistic information makes cross-situational learning more efficient compared with relying on cross-situational statistical or linguistic information alone.

Syntactic bootstrapping (relying on linguistic information) and statistical learning (relying on CSL information) have been viewed as two competing verb acquisition theories. While each account is separately supported by many empirical and modeling studies, different accounts use very different methods. Therefore, it is hard to compare results across experiments to either rule out one account or synthesize them. In the current study, we used the HSP, which has been primarily used to measure the contribution of linguistic information, to provide a unified view of how linguistic information can be integrated in cross-situational learning.

The current study focused on verb learning both within single moments and across multiple moments in a short training session. However, verb learning in the real world takes months and years. Learners' information source is not limited to just statistical and linguistic information. One interesting future study question is to examine whether there are developmental changes in terms of how available linguistic and statistical information is in real-life learning scenarios and how learners utilize other kinds of information over a longer period of time (Frank et al., 2013; Hollich et al., 2000; Monaghan, 2017; Yu and Ballard, 2007). The Emergentist Coalition Model (ECM) offers a unified theory of word learning over development (Hirsh-Pasek et al., 2000). The ECM argues that children have access to a range of information that could help them uncover word meanings. These infor-

mation sources include perceptual, social, and linguistic information. According to ECM, the acquisition of all lexical units is first driven by children's sensitivity to perceptual information to form word-to-world mappings. As children develop, they gradually put more weight on social and linguistic information. To determine how a word is learned, one needs to consider multiple contributing factors and these factors can each independently and collectively impact word learning at different developmental stages.

Acknowledgments

This work is supported by NICHD R01HD074601 and R01HD093792.

References

Arunachalam, S., & Waxman, S. R. (2010). Meaning from syntax: Evidence from 2-year-olds. *Cognition*, *114*(3), 442–446.

Carey, S., & Bartlett, E. (1978). Acquiring a single new word.

Childers, J. B., Bottera, A., & Howard, T. J. (2018). Verbs: Learning how speakers use words to refer to actions.

Childers, J. B., & Paik, J. H. (2009). Korean-and english-speaking children use cross-situational information to learn novel predicate terms. *Journal of Child Language*, *36*(1), 201.

Fisher, C. (1996). Structural limits on verb mapping: The role of analogy in children's interpretations of sentences. *Cognitive psychology*, *31*(1), 41–81.

Frank, M. C., Tenenbaum, J. B., & Fernald, A. (2013). Social and discourse contributions to the determination of reference in cross-situational word learning. *Language Learning and Development*, *9*(1), 1–24.

Gentner, D. (1982). Why nouns are learned before verbs: Linguistic relativity versus natural partitioning. *Center for the Study of Reading Technical Report; no. 257*.

Gillette, J., Gleitman, H., Gleitman, L., & Lederer, A. (1999). Human simulations of vocabulary learning. *Cognition*, *73*(2), 135–176.

Gleitman, L. (1990). The structural sources of verb meanings. *Language acquisition*, *1*(1), 3–55.

Gleitman, L. R., & Gillette, J. (2017). The role of syntax in verb learning. *The handbook of child language*, 413–427.

Goldin-Meadow, S., Seligman, M. E., & Gelman, R. (1976). Language in the two-year old. *Cognition*, *4*(2), 189–202.

Golinkoff, R. M., Jacquet, R. C., Hirsh-Pasek, K., & Nandakumar, R. (1996). Lexical principles may underlie the learning of verbs. *Child development*, *67*(6), 3101–3119.

Halberda, J. (2003). The development of a word-learning strategy. *Cognition*, *87*(1), B23–B34.

Hirsh-Pasek, K., Golinkoff, R. M., & Hollich, G. (2000). An emergentist coalition model for word learning. *Becoming a word learner: A debate on lexical acquisition*, 136–164.

Hollich, G., Hirsh-Pasek, K., & Golinkoff, R. M. (2000). I. what does it take to learn a word? *Monographs of the society for research in child development*, *65*(3).

Horst, J. S., Scott, E. J., & Pollard, J. A. (2010). The role of competition in word learning via referent selection. *Developmental Science*, *13*(5), 706–713.

Maguire, M. J., Hirsh-Pasek, K., & Golinkoff, R. M. (2006). A unified theory of word learning: Putting verb acquisition in context. *Action meets word: How children learn verbs*, 364.

Monaghan, P. (2017). Canalization of language structure from environmental constraints: A computational model of word learning from multiple cues. *Topics in Cognitive Science*, *9*(1), 21–34.

Naigles, L. (1990). Children use syntax to learn verb meanings. *Journal of child language*, *17*(2), 357–374.

Naigles, L. (1996). The use of multiple frames in verb learning via syntactic bootstrapping. *Cognition*, *58*(2), 221–251.

Piccin, T. B., & Waxman, S. R. (2007). Why nouns trump verbs in word learning: New evidence from children and adults in the human simulation paradigm. *Language Learning and Development*, *3*(4), 295–323.

Scott, R. M., & Fisher, C. (2012). 2.5-year-olds use cross-situational consistency to learn verbs under referential uncertainty. *Cognition*, *122*(2), 163–180.

Smith, K., Smith, A. D., & Blythe, R. A. (2011). Cross-situational learning: An experimental study of word-learning mechanisms. *Cognitive Science*, *35*(3), 480–498.

Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, *106*(3), 1558–1568.

Snedeker, J., Gleitman, L. et al. (2004). Why it is hard to label our concepts. *Weaving a lexicon*, *257294*.

Trueswell, J. C., Medina, T. N., Hafri, A., & Gleitman, L. R. (2013). Propose but verify: Fast mapping meets cross-situational word learning. *Cognitive psychology*, *66*(1), 126–156.

Waxman, S. R., & Gelman, S. A. (2009). Early word-learning entails reference, not merely associations. *Trends in cognitive sciences*, *13*(6), 258–263.

Yu, C., & Ballard, D. H. (2007). A unified model of early word learning: Integrating statistical and social cues. *Neurocomputing*, *70*(13-15), 2149–2165.

Yu, C., & Smith, L. B. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological science*, *18*(5), 414–420.

Yuan, S., Fisher, C., & Snedeker, J. (2012). Counting the nouns: Simple structural cues to verb meaning. *Child development*, *83*(4), 1382–1399.

Yurovsky, D., Smith, L. B., & Yu, C. (2013). Statistical word learning at scale: The baby's view is better. *Developmental science*, *16*(6), 959–966.