Seeking Meaning: Examining a Cross-situational Solution to Learn Action Verbs Using Human Simulation Paradigm

Yayun Zhang (yayzhang@indiana.edu)
Andrei Amatuni (aamatuni@indiana.edu)
Ellis Cain (escain@indiana.edu)
Chen Yu (chenyu@indiana.edu)
Department of Psychological & Brain Sciences, Indiana University – Bloomington
1101 East 10th Street, Indiana University, Bloomington, IN, 47405, USA

Abstract

To acquire the meaning of a verb, language learners not only need to find the correct mapping between a specific verb and an action or event in the world, but also infer the underlying relational meaning that the verb encodes. Most verb naming instances in naturalistic contexts are highly ambiguous as many possible actions can be embedded in the same scenario and many possible verbs can be used to describe those actions. To understand whether learners can find the correct verb meaning from referentially ambiguous learning situations, we conducted three experiments using the Human Simulation Paradigm with adult learners. Our results suggest that although finding the right verb meaning from one learning instance is hard, there is a statistical solution to this problem. When provided with multiple verb learning instances all referring to the same verb, learners are able to aggregate information across situations and gradually converge to the correct semantic space. Even in cases where they may not guess the exact target verb, they can still discover the right meaning by guessing a similar verb that is semantically close to the ground truth.

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

Introduction

Children's early vocabularies are composed of overwhelmingly more nouns than verbs (Goldin-Meadow, Seligman, & Gelman, 1976). Many experimental studies on language acquisition have supported the claims that nouns and verbs are learned differently and that verbs are universally more difficult to learn than nouns (Bornstein et al., 2004). One explanation for this verb disadvantage is that to acquire the meaning of a verb, learners not only face the problem of finding the correct mapping between a verb and an action or event in the world, but also the problem of inferring the underlying relational meaning that the verb encodes (Gentner & Boroditsky, 2001; Snedeker & Gleitman, 2004). To do so, verb learners need to discover how their native language combines and lexicalizes many elements of meaning encoded in a single verb (Gentner, 1982).

To give a concrete example of this inherent difficulty in learning action verbs compared with learning concrete nouns: Imagine a parent-child toy play scenario wherein the parent is watching her child play with a toy cube by holding it and turning it around. While watching, the parent says a new word that the child has never heard before. If the new word is an object name, it is easy to infer that the parent is very likely to refer to the cube played by the child at the moment. If the word is an action verb, there are a number of possible verbs that could be used to perfectly describe the situation (e.g. "hold", "play", "show", "twist" and "turn"), any of those actions could be the target referent for the heard verb. For children who do not yet understand verb meanings, verb learning introduces a harder problem compared with noun learning, in that it requires not only finding the correct action-carrying object as a referent, but also uncovering the meanings of the verb in an ambiguous context.

One way to solve the ambiguity problem in early word learning is through cross-situational learning. Several recent studies have shown that both children and adults are good at using cross-situational consistency to figure out the correct word-object mappings (Horst, Scott, & Pollard, 2010; Trueswell, Medina, Hafri & Gleitman, 2013; Smith, Smith, & Blythe, 2011; Smith & Yu, 2008; Yu & Smith, 2007; Zhang, Chen, & Yu, 2019). When language learners see multiple referents and hear multiple words simultaneously, it is not possible for them to learn the mappings between individual words and objects in a single learning instance. However, the correct word-referent mappings will emerge over multiple learning instances as they are likely to co-occur more frequently than incorrect ones. In one study on object-name learning, Smith and Yu (2008) found that 12 to 14-months-olds infants successfully associate object names with their corresponding objects in a cross-situational learning task. In addition, the cross-situational learning solution also seems to apply to verb learning. Childers and Paik (2009) found that 2- to 3-year-old can learn novel verbs by watching multiple visual events with different objects preserving the same action. Scott and Fisher (2012) also showed that 2.5-year-olds are able to use cross-situational statistics to find the correct verb-action mappings. Even though young children can use cross-situational statistics to learn new action verbs in well-controlled experimental contexts, there is evidence suggesting that solving this problem is not so easy in more naturalistic contexts. In contrast to concrete tokens of events used in well-controlled experiments, there is no clear indicator of event boundaries in naturalistic learning situations. Therefore, "packaging" the elements of meanings in the real world can be very challenging, and even adult learners have trouble mapping verbs to actions in those contexts (Gillette, Gleitman, Gleitman, & Lederer,1999).

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 watching a sequence of clips all referring to the same target word. Although participants were given the opportunity to aggregate information cross-situationally, they were only able to guess 45% of the nouns and 15% of the verbs correctly. This finding highlights that naturalistic learning situations can be highly ambiguous, especially for verbs (Gillette, Gleitman, Gleitman, & Lederer, 1999).

Since verb learning situations are inherently ambiguous and many verbs can be used to describe the same situation, inferring the exact verbs mothers in the videos had said can be very challenging. However, this does not rule out the possibility that learners may still gain useful information related to the intended verb meaning from scene observations. We still do not know whether learners are able to integrate visually grounded information from multiple learning scenes and narrow down the semantic space to gradually identify the right verb meaning they need to learn. This is an important theoretical question, as one prerequisite to verb learning is to attend and individuate actions and relations in the environment. In order to uncover verb meanings, learners must first learn to perceive events in the world in ways that align with the concepts embedded in their native language.

To test this idea, we designed a set of experiments using the Human Simulation Paradigm (HSP) originally developed by Gillette et al. (1999). Although adult learners in HSP may not be a perfect model for child learners, understanding how they process statistical information can still provide valuable insights on what statistical information in the environment can be used by young learners for early word learning. To closely approximate the input that young children perceive in everyday learning contexts, we used the video data from the child's first-person view, collected using head-mounted cameras. Recent studies have shown that infants' own egocentric views contain unique properties and distributions that are very different from adults' views, which may be critical for successful learning (Yurovsky, Smith, & Yu, 2013; Bambach, Crandall, Smith, & Yu, 2018).

Three experiments were conducted in the present study. In Experiment 1, 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 was used to select a set of ambiguous instances for Experiments 2 and 3. In Experiment 2, we asked participants to watch a sequence of verb naming instances, all referring to the same target verb, and examined whether the learners could infer the correct verb using aggregated information across multiple instances. Because more than one verb can be potentially used to describe the same scene, our goal for Experiment 3 was to measure the semantic distances of participants' verb choices and then to examine whether learners gradually learn the correct semantic space even though the exact word they chose might not be the target verb.

Experiment 1

Experiment 1 was designed to measure the degrees of ambiguity of a set of verb-naming instances extracted from naturalistic parent-child toy play. We then used those baseline measures to select a subset of ambiguous learning instances as training data for Experiment 2.

Method

Participants Fifty undergraduate students (34 female, M = 19.65 y.o., SD = 1.42) were included. All participants were recruited via university subject pool and received credits for their participation.

Stimuli The video corpus included thirty-two parent-child (child age: M = 19.07 m.o., SD = 3.14, range: 12.3-25.3 m.o.) dyads' play sessions, in which parent-child dyads were asked to play with a set of toys as they naturally would at home for ten minutes (Figure 1). The play interaction was recorded from the child's perspective using a head-mounted camera. From these play interactions, we first transcribed parent speech and then used the transcriptions to identify the moments when parents uttered verbs during play. Among all the verbs mentioned in parent speech, we focused on concrete action verbs with visually grounded verb meanings as they are among the first set of verbs that children learn (Naigles & Hoff, 1998).

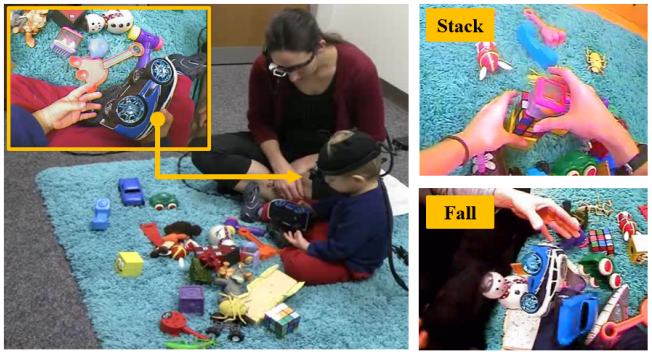

Figure 1. Experimental setup. Parent and child played with a set of toys together in a naturalistic environment. Child's egocentric view video (Upper left) used in the current is captured by the head-mounted camera wore by the child. Screenshots of two verb instances "stack" and "fall" are shown on the right.

Two hundred and ninety-three naming-moment vignettes from the child's view were selected. The target referents include 11 action verbs ("eat", "stack", "knock", "fit", "drive", "cut", "fall", "turn", "put", "hold", "shake"). Each verb had at least twelve naming instances. To avoid the possibility that learners performed cross-situational learning

in this baseline condition, we pseudo-randomized the trials so that both the same target verb and the same dyad would not appear consecutively. As a result, the average number of trials between two same-target trials was 11.56, which made it unlikely for learners to aggregate information across trials.

The original sound of each video was muted, and the verb was replaced by a beep at the onset of the label. All vignettes were 5 seconds long, with the name's onset occurring at exactly the third second. Four additional vignettes with varying difficulties were included as training examples before the experiment to make sure participants understood the task.

Instructions and Procedure

We divided all 293 videos into 3 short 20-min sessions. Participants were instructed to carefully watch some muted short videos of parents playing toys with their children and then guess the intended verb at the moment of parent naming indicated by the beep. They were told to guess concrete action verbs and enter correctly spelled English verbs in the present tense. Each video was only played once, and participants had 20 seconds to enter their best guess after watching each video. No feedback was provided. Among 50 students who participated, 37 did one session, 9 did two and 4 did all three sessions. All participants completed one session within 20 min.

Results & Discussion

Quantifying ambiguity The set of naming vignettes vary in their degrees of ambiguity. As shown in Figure 2A, over 70% of instances are highly ambiguous with less than 40% accuracy. In about 34% of cases, no participants guessed the target verb right. Only in about 3% of instances, almost all participants guessed the target verb right. This result showed that although we only preselected concrete actions that were directly observable, in most cases, guessing the exact verb being uttered was still challenging as participants could come up with many suitable verbs to describe the same perceptual information observed from the video. This finding supports the argument that verb learning is an inherently challenging task.

Trial Selection After measuring each trial's learning accuracy, we defined those trials with less than 40% accuracy as ambiguous trials. Among those trials, we selected 30 trials ($M = 0.13$, $SD = 0.12$) with varying degrees of ambiguity for Experiment 2 (Figure 2B). These 30 trials contain 5 unique target verbs ("knock", "put", "turn", "fall", "hold"), and each verb has 6 different ambiguous vignettes.

Experiment 2

We designed Experiment 2 to examine verb learning when learners were presented with a sequence of ambiguous learning situations all referring to the same target verb. Specifically, we aim to answer two questions: 1) can participants learn the right referent? 2) how likely are they to converge to one referent across trials?

Method

Participants Seventy-three undergraduate students recruited via university subject pool (37 female, $M = 19.57$, $SD = 2.62$) were included in the final sample for data analyses. None of them had participated in Experiment 1.

Stimuli Thirty ambiguous trials (Figure 2B) selected based on baseline measures from Experiment 1. These trials were grouped into 5 blocks with 6 trials in each block, all referring to the same target verb. Two versions were created with different trial and block orders to avoid arbitrary item effects.

Instruction and Procedure Similar to that of Experiment 1, participants were told that they would be trying to guess some verbs by watching blocks of videos of mothers playing with the children. They were aware that all videos within a block were naming the same object and their task was to guess the referred action right after watching each video. Throughout

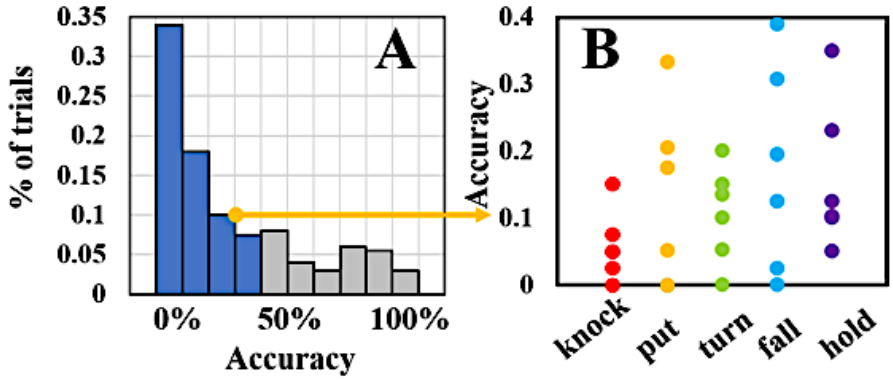

Figure 2. (A) Distribution of guess accuracy across all 293 verb instances. Learning accuracy showed a left skewed distribution. Trials with less than 40% of accuracy (blue) were considered ambiguous. (B) Among those ambiguous trials, 30 (5 referents, 6 instances per referent) were selected for Experiment 2.

the trials, they could change their guess within a block at any given trial. However, if they believed their previous answer was correct, they could choose the same answer again. They were not allowed to go back and change their previous answers and were not given any feedback. After each block, a prompt would appear to remind participants to get ready for the next block of trials.

Results & Discussion

Accuracy of Cross-Situational Learning To quantify learning of the correct target verb, we subtracted the learning accuracies of individual trials with the baseline measures of those trials, which allowed us to directly measure the potential improvement of learning as learners aggregate more information trial by trial. As shown in Figure 3A, we found that participants' first trial improvement was close to zero, which validated our baseline measure. However, this number increased to 11.7% on Trial 4 and to 19.2% on Trial 6. To formally test the improvement over trials, we fit a mixed-effects logistic regression predicting accuracy from trial number with a random effect of subject and version. This model revealed a highly significant main effect of trial number (β =.32, p < .001) over the baseline accuracy (β =5.7, p < .001), indicating significant learning across trials.

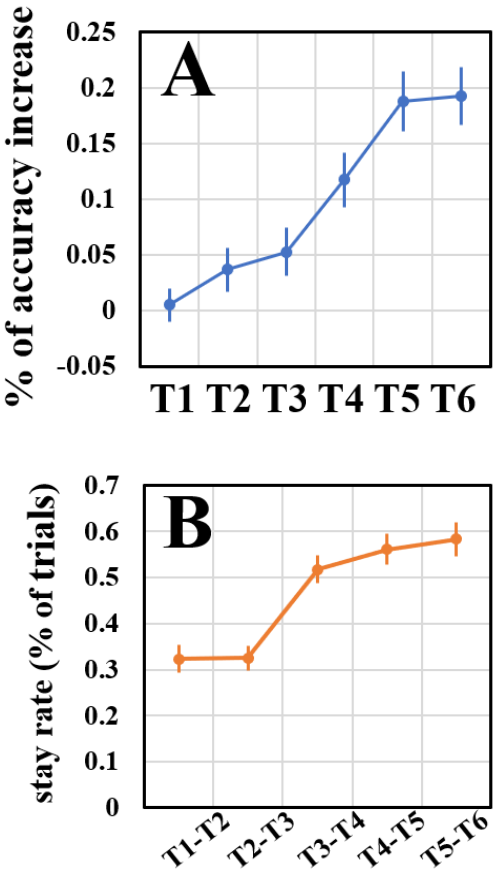

Figure 3. (A) Percentage of accuracy improvement (± 1 SE) across 6 ambiguous naming instances from Experiment 2. (B) Percentage of trials that participants current choice is the same as their previous choice.

Figure 4 shows a concrete trial-by-trial example. In this example, participants saw 6 trials all referring the to the same target verb "put". On Trial 1, only about 12% of learners guessed the correct verb "put" after watching the parent put a snowman on a block, their learning accuracy increased to almost 20% on Trial 3 after watching another naming instance in which the parent put pants on a doll. After watching all six naming instances containing the action "put", learners reached 40% accuracy on the last trial. The dramatic improvement suggests that learners are making progress gradually by integrating what they have seen in previous trials. Even though each pre-selected instance is individually ambiguous (~13% accuracy), all trials together created a much less ambiguous situation for learning.

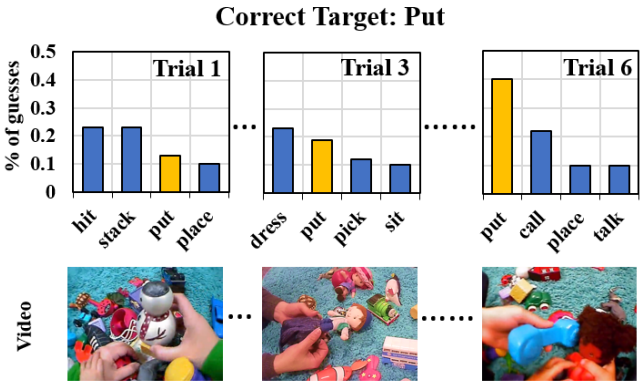

Figure 4. Screenshots from the block of trials with the target word "put". Histograms show the learners' top 4 choices after watching each corresponding video. Yellow bars indicate target verb accuracy.

Convergence of Cross-Situational Learning Target learning result showed that across subjects, more and more learners were able to find the correct target in later trials. However, we still do not know what the learning pattens look like within each subject. How do learners reach the final learning state? To answer this question, we want to measure whether participants converge to one single choice within a block. We counted the number of trials in which participants' choice for the current trial was the same as their previous choice (stay rate) regardless of accuracy. As shown in Figure 3B, on Trial 2, in 32% of cases, participants' choices for the second trial was the same as their previous choices. However, their stay rate increased to 58% for Trial 6, meaning that learners were more likely to keep their previous choice later in a block. To determine whether this difference was statistically significant, we fit a mixed effect model, in which stay rate was coded as -1 if the previous trial was different, 1 if it was the same, and 0 for the participant's first trial. We found that trial order is a significant predictor of stay rate (β =.15, p < .001). This increase of convergence rate suggests that participants are gradually narrowing down their search space to find a verb meaning. By accumulating information

across trials, they became more certain about their guesses and chose to stay with their previous guesses.

Combining accuracy with stay rate, we found that there was an average 38% chance that participants would stay with a wrong choice and the rate of staying with a wrong guess was higher for later trials ($M_{T5\text{-}T6} = 0.50$) compared to earlier trials ($M_{T1\text{-}T2} = 0.28$). Thus, with statistical evidence accumulated over time, learners always converge on a verb regardless of whether it was the target verb. Why did they decide to stay with the non-target verb instead of continuously searching for the target?

One hypothesis is that the converged verb and the target verb may be very closely related. Information provided in the first couple of trials were probably enough for participants to find one possible target verb. At this point, even though this verb is not the target, it is likely to be semantically close to the target. Therefore, learners tended to keep their previous wrong choice because the additional information provided in later trials was not enough to further resolve this ambiguity. In other words, participants were making a *reasonable* mistake by guessing an alternative verb in the same semantic space with the target.

Here is a concrete example supporting our hypothesis from the block of trials with the target word "turn" (Figure 5). After watching Trial 1 video, participants' top four choices for the correct verb were: "twist", "point", "fix" and "turn". On Trial 3, their top four choices became "twist", "turn", "rotate" and "spin". The meaning of "twist" and "turn" started to emerge after 3 trials. On the last trial, about 40% of participants picked "twist" as the target and another 40% of participants picked "turn". It is obvious that "twist" and "turn" are almost identical in terms of the action they are describing in these contexts, and learners are clearly converging to the right semantic space even though for some learners their first choice is not the ground truth target "turn".

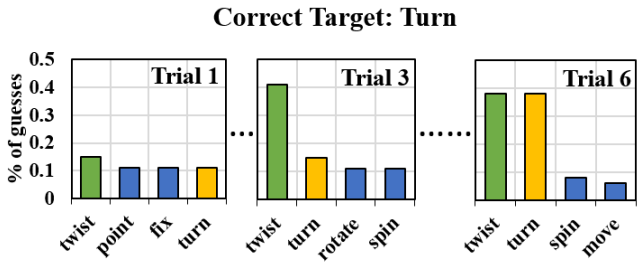

Figure 5. Three sample trials with the target word "turn". Histograms show the learners' top 4 choices after watching each corresponding video. Yellow bars indicate target. Green bars indicate a semantically close verb.

Experiment 3

To test the semantic similarity hypothesis described above, we first need to quantify the semantic distance between target verbs and participants' choices. We tried two commonly used sources of semantic knowledge: WordNet (Wu & Palmer, 1994) and GloVe (Pennington, Socher & Manning, 2014). WordNet is a lexical English database, in which each word is assigned one or more synset, representing different meanings of the word. To measure the semantic distances between our verb pairs, we manually chose target's and response's synsets based on the videos. GloVe embeds words in a vector space where their relative locations are computed based on co-occurrences of words in a given corpus. We used the GloVe model pretrained on 840B tokens of Common Crawl text to create semantic distance measures (Pennington, Socher & Manning, 2014). We discovered that both semantic knowledge bases failed to capture the relationships of action verbs whose similarities can be explicitly observed from videos.

Table 1. Semantic distances between target verb "turn" and four other popular choices from Experiment 2. Distances are ranging from 0-1, low distance (darker shade) indicates high similarity.

Verb Relationships	"Turn" "Twist"	"Turn" "Spin"	"Turn" "Move"	"Turn" "Fix"
WordNet (WUP)*	0.67	0.75	0.6	0.72
GloVe	0.57	0.52	0.33	0.59
Human	0.15	0.22	0.39	0.74

*WordNet scores were converted so small number means high similarity

For example, we measured the semantic distances between target verb "turn" and four other choices from Experiment 2 (Table 1). In both WordNet and GloVe, the word pairs "turn-twist" and "turn-fix" share very similar semantic distances, but based on the perceptually information extracted from videos, "twist" should be much more similar to "turn" than "fix". While these verb relationships can be easily identified using the visually grounded perceptual information extracted from the videos, they are not available in lexical English databases. This is probably because word similarities can be assessed based on many different dimensions (e.g., syntax, semantics, contexts, etc.) and it can be highly dependent on the training corpus. Deriving word representation from text corpora, which integrate rich multimodal properties is an interesting question that is worth further exploration. For our current study, due to the lack of suitable semantic similarity measure, we opt to ask participants to rate the similarities between all possible targets and their responses. Using human rating of verb similarity, we examined whether learners were aggregating information to find the correct semantic space.

Method

Participants Forty-one undergraduate students (30 female, M = 19.22, SD = 0.94) participated in Experiment 3. All participants were recruited via university online subject pool.

Stimuli &Procedure We created a Google form and asked participants to rate the semantic distance of 173 target – response pairs on a 1-7 Likert scale. Seven means the two words were very different (far distance), and 1 means that the two words were almost identical (close distance). Participants

were told that these verbs were from parent speech during naturalistic toy-play. Three additional participants rated all pairwise distances between all possible verb pairs.

Results & Discussion

Semantic Clusters To identify the clusters within the semantic structure of participants' verb choices, we subsampled 40 words (containing three target verbs, "turn", "hold" and "knock") and constructed a 40-by-40 similarity matrix wherein each cell is a pairwise similarity measure among 40 words. Feeding the similarity matrix into Multidimensional scaling (MDS), we visualized the semantic space in a 2D plot (Figure 6) created from the similarities among the 40 words. This visualization, based on quantitative measures, allowed us to see how these verbs were related to each other in the same semantic space. As predicted, "twist", "turn" and "rotate" are closely related in the semantic space. Similarly, "hit", "knock", and "break" are also closely related. Verbs like "play", "change", "move" are shared among clusters, suggesting that they may contain more generic meanings and can be broadly applied to a play context.

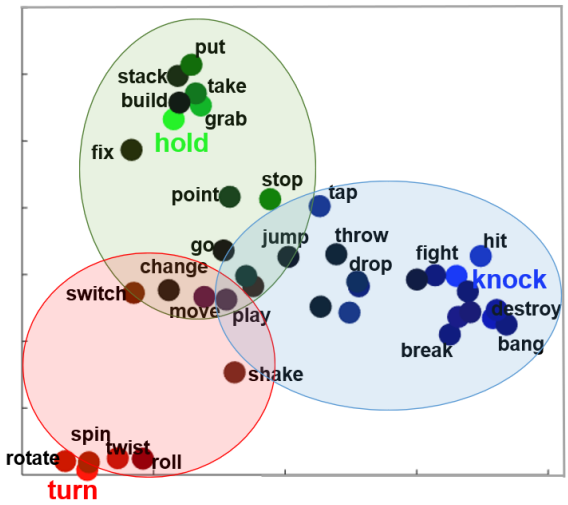

Figure 6. MDS plot showing semantic distances of 40 verbs. Three target verbs "turn", "hold", "knock" are marked in red, green and blue. Each word's dot color is determined by three channels of the RGB values. The three values are weighted semantic distance between the word itself to "turn" (R), "hold" (G), "knock" (B) respectively. Similar colors and closer clustering indicate small semantic distances.

Statistical Learning of Verb Meaning We also used MDS plots to show the convergence of verb meaning. Using block "turn" as an example (Figure7). On Trial 1, learners' guesses were distributed more diffusely and very few people located the right semantic space early on. However, learners started to shift towards the right area by either guessing the right target "turn" or guessing words like "spin", "twist", "rotate", which all share similar meaning to "turn". On the last trial, we can see a clear convergence that the majority of participants have found the correct verb meaning. Learners cannot further distinguish "turn" and "twist" because the visual information extracted from the current videos alone is not enough to disambiguate the meaning of the two closely related words.

To further quantify whether participants were learning the correct verb meaning by converging to the correct semantic space across trials. We measured how semantic distances between target and response changed across trials. As shown in Table 2, on Trial 1, about 65% of participants' guesses were far away from the target (>4 distance on a 1-7 scale) and only 8% of participants guessed the correct target. As learners accumulate more information from additional trials, the distances between target and response gradually become smaller. On Trial 6, about 28% of participants guessed the correct target and 35% got close (<4 distance) to the correct verb meaning. We fit a mixed effect model predicting semantic distance from trial number and we found the model to be statistically significant (β =.32, p < .001).

Table 2. Semantic distances between target and response. Numbers indicate percentage of instances.

Distance / Trial	0 on target	1-4 close	4-7 far
1	0.08	0.27	0.65
2	0.17	0.24	0.59
3	0.23	0.34	0.43
4	0.30	0.34	0.36
5	0.31	0.28	0.41
6	0.28	0.35	0.37

Figure 7. MDS plots showing semantic distances for 3 "turn" trials. Dot size indicate proportion of guesses. Dot shades indicate distance to target.

General Discussion

In this study, we conducted three experiments using HSP and found that learning the meaning of concrete action verbs is challenging because learners can extract many different meanings from watching the same concrete action (Experiment 1). However, when shown multiple learning instances referring to the same verb, learners can gradually discover the correct verb meaning by aggregating information across trials (Experiment 2). In some cases, even though learners fail to guess the exact target verb at the end, they are still converging to a verb that is semantically close to the target, indicating that they are still using the information accumulated through early trials to locate the correct semantic space (Experiment 3). Learners are moving towards the correct semantic space by integrating statistical information. Although verb learning is challenging, our findings suggest a cross-situational solution to solve this problem.

Our findings also provided insights regarding why verbs are harder to learn than nouns. Word learning is essentially a multimodal mapping problem (Quine's *gavagai* problem, 1960). Learners need to use pieces of information from different sensory modalities to build a shared conceptual system. Nouns are easier to learn because they tend to be more concrete. In other words, they have more distinctive structural relationships that are easier for conceptual alignment (Roads & Love, 2020). However, verbs are more complicated. Even for concrete action verbs that are directly observable, they tend to contain conceptually ambiguous and overlapping spaces. With limited learning data, when all concepts are equally similar ("turn" and "twist" is a good example), there is no structure in the perceptual similarity relationships that can further resolve this ambiguity at the moment. However, word learning is a continuous process where, although learners may make sensible mistakes at the moment, they can further refine and distinguish verb meanings using new information gathered from the learning environment.

Acknowledgments

This work is supported by National Institute of Child Health and Human Development R01HD074601 & R01HD093792.

References

Bambach, S., Crandall, D., Smith, L., & Yu, C. (2018). Toddler-inspired visual object learning. In *Advances in neural information processing systems* (pp. 1201-1210).

Bornstein, M. H., Cote, L. R., Maital, S., Painter, K., Park, S. Y., Pascual, L., ... & Vyt, A. (2004). Cross-linguistic analysis of vocabulary in young children: Spanish, Dutch, French, Hebrew, Italian, Korean, and American English. *Child development*, *75*, 1115-1139.

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*, 201-224.

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

Gentner, D., & Boroditsky, L. (2001). Individuation, relativity, and early word learning. *Language acquisition and conceptual development*, *3*, 215-256.

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

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

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

Naigles, L. R., & Hoff-Ginsberg, E. (1998). Why are some verbs learned before other verbs? Effects of input frequency and structure on children's early verb use. *Journal of child language*, *25*, 95-120.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532-1543.

Quine, W. V. O. (1960). Word and object Cambridge. *MA: MITPress*.

Roads, B. D., & Love, B. C. (2020). Learning as the unsupervised alignment of conceptual systems. *Nature Machine Intelligence*, 1-7.

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

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

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

Snedeker, J., & Gleitman, L. (2004). Why it is hard to label our concepts. *Weaving a lexicon*, 257-294.

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*, 126-156.

Wu, Z., & Palmer, M. (1994). Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics,* 133-138.

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

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

Zhang, Y., Chen, C. H., & Yu, C. (2019). Mechanisms of Cross-situational Learning: Behavioral and Computational Evidence. *Advances in child development and behavior*, *56*, 37-63.