Grounding Action Verbs in Egocentric Visual Perception

Yayun Zhang[1], Ellis Cain[2], David Crandall[3], Chen Yu[1]

yayunzhang@utexas.edu, ecain@ucmerced.edu, djcran@iu.edu, chen.yu@austin.utexas.edu

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

Abstract

It has been conjectured that verb learning is hard because verb meanings are not readily "packaged" from the physical world. To provide new empirical evidence on this account, we analyzed egocentric video collected from natural toy-play interaction and focused on the naming events when action verbs were uttered in parent speech. Using the Human Simulation Paradigm, we showed egocentric videos of those naming events to adult observers and asked them to guess the target verb in parent speech. We found that adult observers used many different verbs to describe the same visual event, and only one of them matched with the verb in parent speech. We analyzed mismatched verbs and identified several sources of mismatch, and found that all of the mismatched verbs are relevant to the target verb, but they capture different properties (temporal, semantic, etc) of the visual event. We also found that different naming events for the same verb also differ in terms of the degree of ambiguity. Taken together, the results in the present paper provided new evidence from the child's view, showing that verb learning is hard not only because multiple possible meanings are embedded in each learning situation, but also because these candidate meanings expand across multiple dimensions of the physical world, overlap with each other, and relate to the target meaning in many different ways.

Keywords: word learning, verb learning, Human Simulation Paradigm

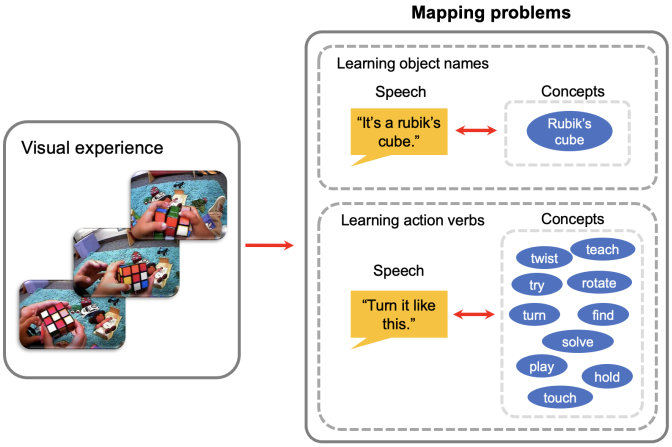

Figure 1: Compared with learning object names, inferring the precise meaning of a verb is challenging because there are many perceivable concepts that can be extracted from a visual event.

Introduction

In the word learning literature, there is consensus that verb-meaning mappings are harder to learn than noun-object mappings (Gentner, 1982; Goldin-Meadow et al., 1976), and that the cognitive processes required to learn verbs may differ from the processes of learning nouns (Gentner and Boroditsky, 2001; Golinkoff and Hirsh-Pasek, 2006; Imai et al., 2005). However, there is no agreement on why verbs are harder to learn. Multiple accounts have been proposed to explain the verb learning challenge from different perspectives. For example, some theorists argued that verbs are harder to learn because verbs are ephemeral and transient in nature, whereas nouns are in general more stable (Slobin, 2001). Others suggested that verbs are harder to learn because verbs do not co-occur in time with corresponding actions but most often precede the actions (Tomasello and Kruger, 1992). Moreover, one of widely accepted accounts by Gentner (1982) states that the meaning of a verb is not readily packaged and conceptually hard to infer from the non-linguistic context.

To take a concrete example, imagine a mother is playing a Rubik's cube with her child (Figure 1). At this moment, if the child hears the mother say an object name that the child has never heard before (i.e., "It's a Rubik's cube"), then it is relatively straightforward for the child to assume that the Rubik's cube is the intended referent by the parent because Rubik's cube is the only object in view and learners tend to assume a noun refers to a whole object (Markman, 1989). However, if the child hears an utterance containing a novel verb (i.e., "Turn it like this!"), there are multiple candidate concepts that can be extracted from the learning moment and serve as the potential referents of that verb: some concepts provide a general description of the visual event (i.e., play, move); some describe a specific action at the moment (i.e., turn, rotate); and others are abstract concepts that can be inferred from the visual event (i.e., solve, try). Unlike noun learning wherein learners can use heuristics, such as the whole-object assumption, to narrow down candidate referents in consideration, there seems to be few assumption or bias that can be utilized to infer the meaning of a verb. For young learners who do not yet have perfect conceptual understanding of the world, they need to "discover" the referents themselves. As Gentner noted, "Referents are not simply 'out there' in the experiential world", and that the key challenge for verb learning is to discover what properties of the observed event are related to the meaning associated with a heard label (Gentner, 1982).

Table 1: Descriptive statistics for selected verbs

	Light verbs	Heavy verbs
Total number of verbs	16	24
Total number instances	400	369
Average number of instance per verb	25	15.38
Average number of responses per instance	6.04	5.99
Sample verbs	go, get, play, make, take	turn, shake, stack, cut, touch

Gentner's account has been supported by empirical evidence. Gillette, Gleitman, Gleitman, and Lederer (1999) conducted the seminal "human simulation" study by asking adults to guess the words mothers used after watching muted videos recorded during parent-child interaction. In the study, adults correctly guessed the missing nouns in 45% of the cases, and the missing verbs for only 15% of the cases. While the accuracy of mental verbs (e.g. "think") is close to 0%, the accuracy for more concrete verbs (e.g. "shake") is also very low (Gillette et al., 1999). These findings provided evidence showing that 1) finding the correct verb-referent mapping is a challenging task even for adult learners who have perfect conceptual knowledge of the world; and 2) there is variability among different types of verbs in that concrete verbs are more learnable than abstract ones.

In literature, there is no agreement on a clear theoretical definition of concrete vs. abstract verbs. One metric used to quantify this property is to measure the number of co-occurring objects that are associated with a particular verb (Maouene et al., 2011; Theakston et al., 2004). Using this metric, verbs can be put into "light" and "heavy" categories. Light verbs, such as do, make, get, take, and go, are more abstract in that it can co-occur with different objects to label a wide range of events (e.g. making a coffee and making bed) wherein they may have little in common. Heavy verbs, such as spin, hammer, hold, and read, are more concrete and specific to a particular action involving one or a small number of objects.

The present study focused on two groups of concrete verbs (one light group, one heavy group) that have visually grounded meanings. Even for those concrete verbs, when hearing them, there are usually more than one visual events present, and each visual event can be described using more than one verbs. Our overarching hypothesis is that even for concrete verbs, multiple concepts may be extracted from an observed visual event and those concepts may be considered as candidate meanings for the heard verb. The goal of the present study is to quantify what types of concepts may be extracted from a visual event and the relations between those candidate concepts and the meaning of the target verb.

Toward this goal, we used the Human Simulation Paradigm (HSP) in which adult observers provide a verb response to describe an observed visual event. Different verb responses of the same visual event are taken together to quantify the degree of ambiguity at the naming moment. Different from the original HSP studies, the video stimuli in the present study were taken from the child's egocentric view recorded during naturalistic parent-child toy play (Pereira et al., 2014). It has been shown that in everyday contexts such as toy play, the child's view is markedly different with the adult's view or a third-person view (Yurovsky et al., 2013). Because the goal is to measure what concepts may be perceived and extracted from a visual event and thereafter considered as the meaning of a verb, using egocentric view videos captures the visual information available to the learner at the naming moment.

We hypothesized that even with concrete verbs with perceptually grounded meanings, multiple possible meanings can be extracted from a visual event, which makes verb learning inherently challenging. Further, we examined the relations among those concepts embedded in a visual event and categorize them to identify several different sources that create ambiguity for verb learning. Built upon Gentner's conceptual framework, the present paper provided new evidence from the child's view, showing that verb learning is hard not only because multiple concepts are embedded in each learning situation, but also because these candidate concepts expand across multiple dimensions of the physical word, overlap with each other, and relate to the target meaning in many different ways.

Method

Participant

Two hundred and seventy-eight participants (112 females, 114 males, 52 did not report, age: M = 44.2 years old, SD = 11.3) recruited through Amazon Mechanical Turk completed the study.

Stimuli

The video corpus included 56 parent-child (child age: M = 16.75 m.o., SD = 4.84 m.o., range: 11.8 - 25.3 m.o.) play sessions. The dyads were told to play with a set of toys as naturally as they would do at home for 10 minutes. Two toy sets (24 toys and 28 toys) were used in the current study. Both toy sets included common toy objects like vehicles, animals, tools, food. The play session was recorded from the child's perspective using head-mounted cameras. We transcribed parent speech and then identified verb naming moments from the full speech transcription. We only coded action verbs that can be visually grounded in a visual event.

Therefore, attention-getting verbs (i.e. look, see) and auxiliary verbs (i.e., be, do, can, may, must) are not included. In total, we coded 1423 verb naming instances for 145 unique verbs. The top 10 most frequent verbs in our corpus are: stack, go, put, make, play, get, shake, come, twist, drive, knock. We then extracted egocentric video clips around these verb naming moments. All clips were 5 seconds long, with the naming onset occurring at exactly three seconds. Following the HSP, 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 were created as training examples and were presented to participants before the start of the experiment to ensure that participants understood the task.

Instructions and Procedure

Participants watched a set of short egocentric videos recorded from parent-child toy play. They were told that these videos contain naming moments when the parent produced a spoken utterance containing an action verb. Because the original sound of the videos was removed, participants need to guess the intended verb indicated by the beep. Each video was played once, and participants had 20 seconds to enter a verb in its present tense without getting any feedback. Each participant went through fifty trials randomly selected from the entire corpus. The whole session lasted about 20 minutes.

Results

For subsequent data analyses, we only included verbs with at least 7 unique trials, and trials with at least 5 responses from different participants. With those selection criteria, 40 unique verbs were included. We grouped these 40 verbs into heavy and light categories based on two categorization schemes provided in Theakston et al. (2004). Eight out of 40 verbs we chose were not on their non-exhaustive list. Therefore, we coded them using our best judgement. For example, for verbs (i.e., hammer, rake, saw) that were specifically related to a particular toy, we coded them as heavy verbs. Descriptive statistics regarding these two categories are listed in Table 1.

Number of unique meanings

We first quantified concepts embedded in each naming instance by calculating the number of unique verbs provided by participants. On average, participants identified 4 unique verb meanings (M = 3.88, SD = 1.49) per trial. In more than 95% of trials, they extracted more than one verb meaning from the same visual event. This pattern held true for both heavy and light verbs (Figure 2). We did not find a significant difference between the two groups (Light verbs: M = 4.11, SD = 1.46, Heavy verbs: M = 3.63, SD = 1.48, $t(38)$ = 1.2, ns), suggesting that for each target verb in both groups, multiple candidate meanings extracted from a visual event are in consideration.

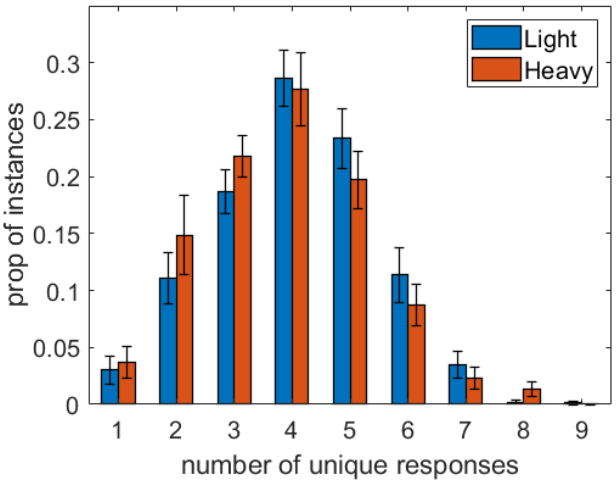

Figure 2: Unique number of responses

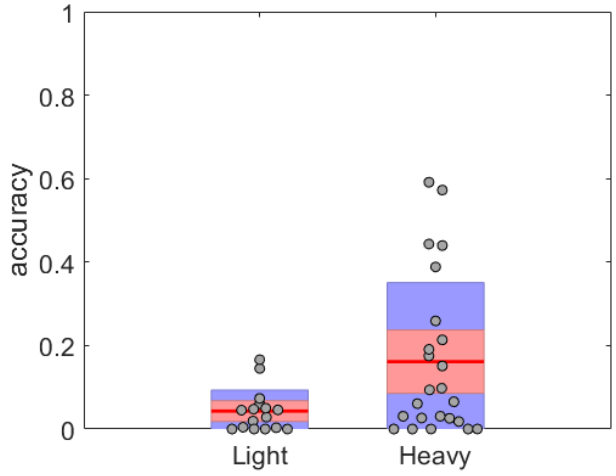

Figure 3: Mean accuracy of HSP responses

Accuracy of responses

Given multiple verbs from each visual event, we next examined how well those verbs matched the target verb in parent speech. We found that the accuracy of identifying the correct verb meaning is low (M=11%, SD = 0.16). Further, adult observers were more likely to correctly identify heavy verbs (M=16%, SD = 0.19) than light verbs (M=4.3%, SD = 0.05); $t(38)$ = 2.43, p = 0.02, Figure 3).

Given that learning accuracy is low on average, it is not clear whether there are item-level differences. Is it the case that some items are easier to learn than other or all items are uniformly difficult? To take a closer look at the item-level accuracy, we plotted accuracy distribution for each verb and ranked ordered them based on their mean accuracy scores. As shown in Figure 4, x-axis represents accuracy and y-axis represents proportion of instances. Hotter color indicates more responses fall in the corresponding accuracy range. For ex-

ample, for the verb "saw", about 35% of instances fall between 0 to 10% accuracy range, 10% of instances fall between 10 to 20% accuracy range, 20% of instances fall between 60 to 70% accuracy range and another 35% of instances fall between 80 to 90% accuracy range. We found that most verb-learning instances fall on the low accuracy end of the scale, meaning that the majority of verb learning instances are quite challenging. Among the small number of items with relatively high accuracy scores, the majority of them are heavy verb instances. There is a general trend that heavy verbs tend to have higher accuracy scores and wider accuracy distributions than light verbs.

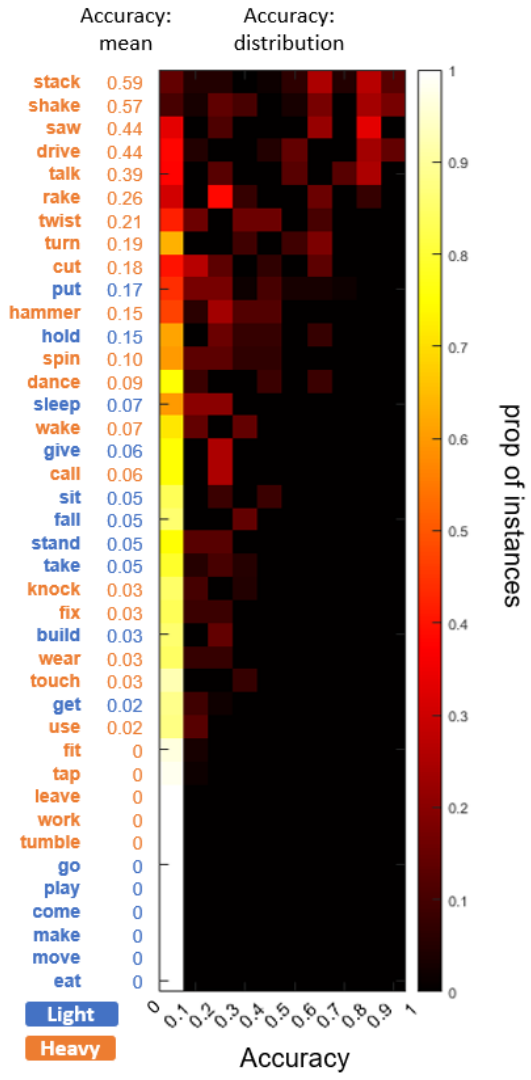

Figure 4: Accuracy distribution for each of the 40 verbs. Verb are ranked by mean frequency and color-coded to indicate light and heavy verbs.

In sum, both group and item-level accuracy measures indicate that most verb-learning instances are very difficult. With many candidate verbs identified from a visual event, only one of them matched with the target verb in parent speech. Further, it is more difficult to find the meaning for light verbs than heavy verbs, suggesting a higher degree of ambiguity for light verbs.

Sources of mismatch

Given a visual event, the adult observers identified several verbs and only one of them was the target verb in parent speech. Meanwhile, all of the other verbs are also inferred from a visual event and therefore their corresponding meanings may be considered to map to the target verb. Thus, analyzing those mismatched verbs allowed us to examine what types of concepts may be extracted from a visual event and the relations between those mismatched concepts and the meaning of the target verb.

Provided with a list of mismatched verbs for a visual event, we identified five relation types between candidate verbs and the target verb: `subordinate`, `superordinate`, `concurrent`, `sequential` and `synonymous`, and used the following coding scheme to group individual verbs into one of the five mismatched categories:

- `subordinate`: the target verb describes a general activity (e.g. "play") while a candidate verb refers to a specific action (e.g. "put") within the activity.
- `superordinate`: the target verb refers to a concrete action (e.g. "put") while a candidate verb describes a general activity such as "play".
- `concurrent`: when there are multiple actions co-occurring in time, a candidate verb is used to describe one of the actions but not the target one mentioned in parent speech. For example, the child was shaking a toy helmet using his left hand while throwing a toy block away using his right hand. A candidate verb from adult observers is "shake" while the target verb in parent speech is "throw".
- `sequential`: when there are multiple actions occurring in a sequence, a candidate verb is used to describe a segment of the whole sequence that is not the target segment described in parent speech. For example, when the child reached for a toy and grabbed it, an observer described the event as "grab" while "reach" was mentioned in parent speech.
- `synonymous`: a candidate verb is synonymous to the target verb. For example, both "twist" and "turn" can be used interchangeably to describe the visual event in which a child was playing with a Rubik's cube.

Following this coding scheme, three trained coders independently annotated 1306 unique target-response pairs and reached 85% agreement. We then assigned a category for each mismatched verb. Using the coded data, we aimed at answering two questions: 1) are mismatched verbs of the same visual event from different categories? If so, are there differences of these mismatched categories over different verbs and between heavy and light verbs? 2) Are there verb type

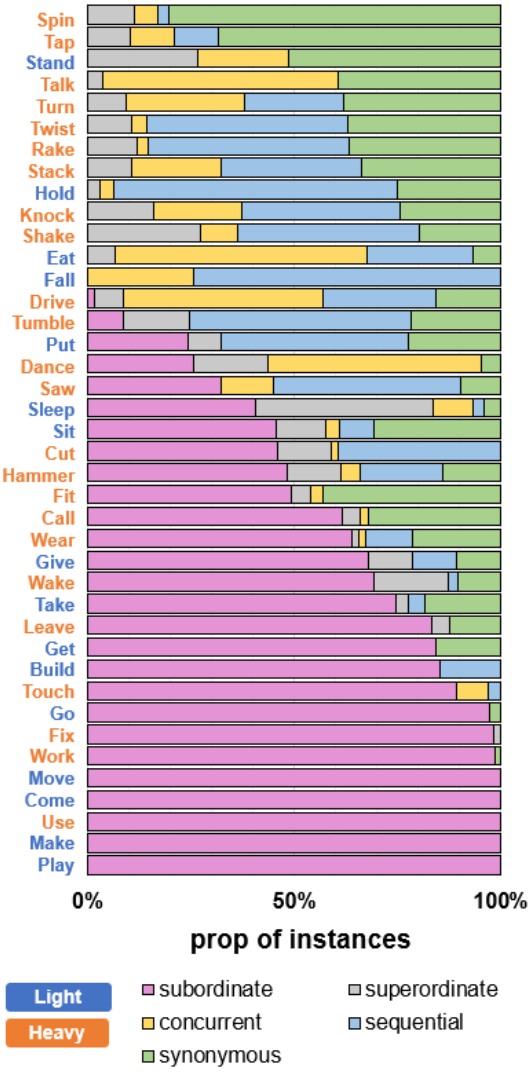

Figure 5: Distributions of mismatched categories for all target verbs.

differences at the individual instance level? In other words, do different learning instances of the same verb share similar or different mismatched types?

To answer the first question, we measured how often the 5 mismatched types occur for each of the 40 verbs. Most verbs contain mismatched verbs that belong to multiple types. For example, for the target verb "spin", the mismatched verbs, such as "move", "hold","push", and "rotate", belong to 4 different types (`superordinate`, `concurrent`, `sequential` and `synonymous`, etc.) This response pattern suggests that different types of mismatched meanings are available for visual events of the same target verb and these meanings are all relevant to some aspects of the event and share some properties with the intended target meaning. From a verb learner's perspective, the challenge is to not only figure out the correct verb from many candidates within one dimension, but also find the shared elements across multiple dimensions (event,

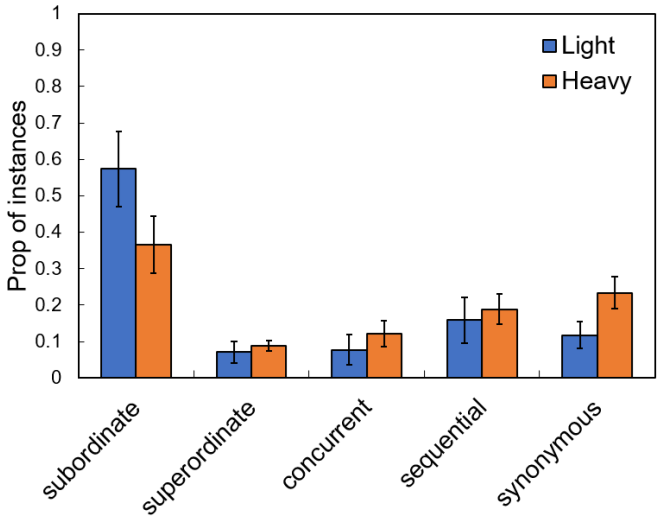

Figure 6: Comparison between heavy and light verbs across the 5 mismatch categories.

action, semantics, etc). Because different meanings derived from multiple dimensions of a visual event can all be candidate referents for the heard verb - there is no easy way to "package" elements of a visual event into a single lexical unit.

Second, there is a difference between heavy and light verbs (Figure 6). Subordinate type in which learners' responses can be viewed as a component of the target is more common in light verbs, whereas concurrent, sequential and synonymous categories tend to be more common in heavy verbs. This pattern suggests that light and heavy words may be hard to learn in different ways. For heavy verbs, the key is to identify common features across multiple overlapping dimensions. For light verbs, the key is to inform a more general and more abstract meaning from many concrete events. The verb-level differences we observed here could be viewed as one of the explanations of why some verbs are learned before other verbs (Naigles and Hoff-Ginsberg, 1998).

Although for each verb, observers' responses fall in multiple mismatched categories, it is not clear whether this variability is coming from the same or different learning instances of that particular verb. To find out whether learners' responses consistently fall in one mismatched category or different mismatched categories at the instance level, we extracted all category types across different responses for the same learning instance and counted how many categories each instance had. We found that on average, at the individual instance level, observers' responses belong to two categories (M = 1.70, SD = 0.46). This pattern is consistent for heavy (M = 1.76, SD = 0.42) and light instances (M = 1.61, SD = 0.50, Figure 7), suggesting that even for the same verb, multiple training instances can be quite different. This instance-level differences can be be viewed as another factor contributing to the verb learning challenge.

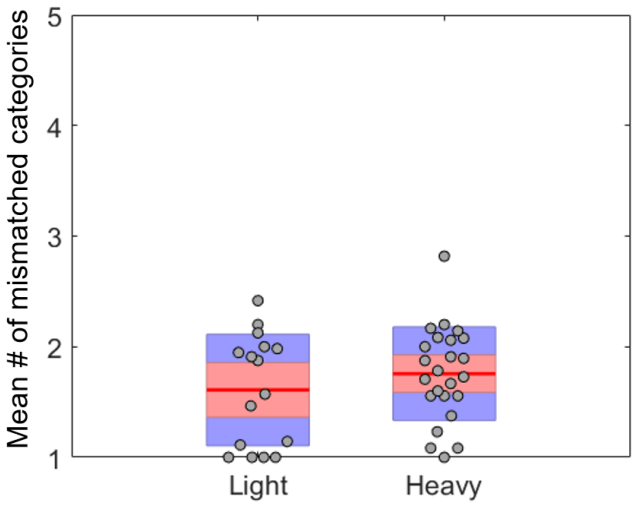

Figure 7: Mean number of mismatched categories per instance.

Discussion

The current study extended our existing knowledge on verb learning by showing that verb learning is challenging not only because multiple concepts are embedded in each learning situation, but also because these candidate concepts expand across multiple dimensions of the physical world. To successfully extract the precise meaning of a verb, learners need to do inductive learning by identifying what properties or rules are relevant to the verb meaning from many overlapping dimensions.

Because there are always multiple verb concepts embedded in each learning situation, it is nearly impossible to narrow down the relevant properties using information from one verb-learning event alone. Luckily, in real life, young children's learning environment is highly structured, yielding utterances that follow regularities. For example, parents rarely label an object only once and move on to labeling another object in the next sentence. Instead, they tend to form extended episodes of discourse about one object followed by a period when the object is rarely mentioned, then parents may come back to talk about it again in another conversation later on (Frank et al., 2013; Suanda et al., 2016). Given the dynamic structures of word learning environment, one interesting question we can ask is how infants make progress gradually by integrating what they have learned before? Can learners extract meanings from multiple verb-learning instances?

Many researchers have used Cross-Situational Learning (CSL) tasks to study the word learning process under uncertainty. Although learners may not be able to identify the correct word-referent mapping on a single exposure, if learners can combine information across multiple exposures, they are able to determine the most probable referent by integrating multiple candidate sets over time (Smith and Yu, 2008; Trueswell et al., 2013; Yu and Smith, 2007; Zhang et al., 2021). The cross-situational learning solution has also been studied in verb learning. Previous work has shown that 2- to 3-year-old can learn novel verbs by watching multiple visual events with different objects preserving the same action (Childers and Paik, 2009; Scott and Fisher, 2012). However, one limitation of the existing verb CSL design was that the visual referent provided in those studies were all concrete token of events, meaning that the visual referents have already been packaged with clear event boundaries. Although these concrete tokens of events can be viewed as one source of ambiguity embedded in naturalistic verb-learning contexts, it does not seem to capture the multidimensional nature we observed from verb-learning events in the real world.

Our findings provide new insights regarding why verbs are hard to learn. Even for concrete action verbs that are directly observable from the child's view, finding the precise meaning of them is not a trivial task because learners need to identify relevant properties both within and across multiple overlapping dimensions of the real world. Verb learning is likely to be a prolonged process where learners have to continuously refine a verb concept through many different encounters. Although learners may not get to the precise meaning of a verb at the moment, the multiple concepts they perceive from each learning instance are still critical building blocks that can help them form an accurate verb concept down the road.

Acknowledgments

This work is supported by NICHD R01HD093792.

References

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.

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*.

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*(2), 135–176.

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

Golinkoff, R. M., & Hirsh-Pasek, K. (2006). Introduction: Progress on the verb learning front. *Action meets word: How children learn verbs*, 3–28.

Imai, M., Haryu, E., & Okada, H. (2005). Mapping novel nouns and verbs onto dynamic action events: Are verb meanings easier to learn than noun meanings

for japanese children? *Child development*, *76*(2), 340–355.

Maouene, J., Laakso, A., & Smith, L. B. (2011). Object associations of early-learned light and heavy english verbs. *First Language*, *31*(1), 109–132.

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

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*(1), 95–120.

Pereira, A. F., Smith, L. B., & Yu, C. (2014). A bottom-up view of toddler word learning. *Psychonomic bulletin & review*, *21*(1), 178–185.

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.

Slobin, D. I. (2001). Form/function relations: How do children find out what they are?

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

Suanda, S. H., Smith, L. B., & Yu, C. (2016). The multisensory nature of verbal discourse in parent–toddler interactions. *Developmental neuropsychology*, *41*(5-8), 324–341.

Theakston, A. L., Lieven, E. V., Pine, J. M., & Rowland, C. F. (2004). Semantic generality, input frequency and the acquisition of syntax. *Journal of child language*, *31*(1), 61–99.

Tomasello, M., & Kruger, A. C. (1992). Joint attention on actions: Acquiring verbs in ostensive and non-ostensive contexts. *Journal of child language*, *19*(2), 311–333.

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.

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

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.

Zhang, Y., Yurovsky, D., & Yu, C. (2021). Cross-situational learning from ambiguous egocentric input is a continuous process: Evidence using the human simulation paradigm. *Cognitive science*, *45*(7), e13010.