Why Some Verbs are Harder to Learn than Others – A Micro-Level Analysis of Everyday Learning Contexts for Early Verb Learning

Siyun Liu (liusy@mail.ccnu.edu.cn)
Key Laboratory of Adolescent Cyberpsychology and Behavior, Ministry of Education
School of Psychology, Central China Normal University
No.152 Luo Yu Road, Wuhan, Hubei, 430079, China

Yayun Zhang (yayzhang@indiana.edu)
Chen Yu (chenyu@indiana.edu)
Department of Psychological and Brain Sciences, Indiana University- Bloomington
1101 E. 10th Street, Bloomington, IN, 47405, USA

Abstract

Verb learning is important for young children. While most previous research has focused on linguistic and conceptual challenges in early verb learning (e.g. Gentner, 1982, 2006), the present paper examined early verb learning at the attentional level and quantified the input for early verb learning by measuring verb-action co-occurrence statistics in parent-child interaction from the learner's perspective. To do so, we used head-mounted eye tracking to record fine-grained multimodal behaviors during parent-infant joint play, and analyzed parent speech, parent and infant action, and infant attention at the moments when parents produced verb labels. Our results show great variability across different action verbs, in terms of frequency of verb utterances, frequency of corresponding actions related to verb meanings, and infants' attention to verbs and actions, which provide new insights on why some verbs are harder to learn than others.

Keywords: verb learning, motion verb, attention, head-mounted eye-tracking, infant-parent dyads

Introduction

Language learning depends on both the internal learning mechanisms and the data on which those mechanisms operate. Many experimental studies have focused on examining the internal learning mechanisms by using well-controlled and well-balanced stimuli as the input. A recent trend in the field of language acquisition is to examine natural statistics in everyday learning contexts (e.g. Pereira, Smith, & Yu, 2014). For example, recent studies have shown that both the quantity and quality of parent language input are predictive of children's later language development (Hart & Risley, 1995; Hoff, 2003; Weisleder & Fernald, 2013). In the present study, we used the same approach to examine the input for early word learning. One of the challenges in early word learning is to figure out the correct mapping between a word and a referent (Quine, 1960). Given many possible referents in the moment when a word is heard, young learners need to attend to the right referent at the right time in order to learn the meaning of a word. However, we do not yet know what input from the environment is available to the child and what input attended by the child is therefore processed by the internal learning mechanisms.

A large proportion of early vocabulary is composed of concrete nouns and concrete verbs. Previous studies on learning concrete nouns found that children need to select and attend to the right object at the right time from an ambiguous learning environment when hearing its name (Golinkoff, Hirsh-Pasek, Bailey, & Wenger, 1992; Yurovsky, Smith, & Yu, 2013; Yu & Smith, 2007; Gleitman & Trueswell, 2018). In addition, Pereira, Smith and Yu (2014) found that when the named target is visually large and more centered in the child's view and when these optimal visual properties last longer before and after parent's naming, children are more likely to attend to the named object and learn its label. Thus, learning object names with perceptually grounded meanings requires not only hearing the words from parent speech but also showing sustained attention to the intended referent. However, little is known about whether learning concrete verbs also requires young learners' sustained attention when mapping verbs to visually grounded actions. Most experimental studies on verb learning have been focused on testing how well young children build verb-action mappings when presented with a verb and an action in well-controlled laboratory settings (Imai et al., 2008; Hirsh-Pasek & Golinkoff, 1996; Maguire et al., 2008; Golinkoff et al., 2002; Pulverman, Golinkoff, Hirsh-Pasek, & Buresh, 2008; Monaghan, Mattock, Davies, & Smith, 2015, Messenger, Yuan, & Fisher, 2015; Scott & Fisher, 2012). The learning tasks for young children in those experimental setups were well-controlled to minimize distraction, which is very different from learning verbs in the real world. Referential uncertainty created during naturalistic interactions may be different from that created for traditional lab tasks, thus it may influence how children process information differently.

Imagine a naturalistic context for early verb learning such as toy play, when a parent names a verb (e.g. "Can you shake it?") while demonstrating the shaking action. The meaning of "shake" is presented briefly as the parent is not likely to keep shaking the object. If the infant does not attend to the action when hearing the word "shake" and when the action is produced, it would be impossible for the infant to build the association between the word "shake" and the action "shake". This example reflects the transient nature of the action referent and lead to important research questions related to early verb learning that have not been examined at the

perceptual and attention levels. For example, compared with object names, how frequently do parents mention action verbs in their speech in everyday learning contexts? When parents produce a verb in speech, how likely there is a corresponding action in the learning environment that reveals the meaning of the verb? If there is an action in accompany with parent speech, how likely do infants attend to the action to build a verb-action mapping?

To answer these questions, we need to examine parents' and children's behaviors from natural learning environments. We used head-mounted eye-tracking techniques to record fine-grained multimodal behaviors during parent-infant joint play. We analyzed parent speech, parent and infant action, and infant attention at the moments when parents produced verb labels. By doing so, we will be able to provide new evidence on how easy or hard for young children to learn early verbs and discover new elements -- at the attentional level -- that matter to early verb learning. Our overarching goal was to quantify word-referent co-occurrence statistics in parent-child interaction from the learner's perspective and examine what information infants select to attend when a verb is heard.

Method

Participants

Thirty-three infant-parent dyads with infants (12 female) ranging from 15.2 to 25.3 months (M = 19.52, SD = 2.42) were included in the final sample.

Stimuli and Experimental Setup

Parents and their infants were invited to play with a set of 24 toys in a playroom (Figure 1A). The toys were randomly spread out across the floor at the beginning of each play session. Parents and infants both sat on the floor and parents were told to sit in any orientation with their child but were instructed to try to keep their child sitting on the ground as much as possible during the play session. We observed that parents and infants naturally generated various types of manual actions during toy play. For example, they used a toy saw to pretend to cut other objects; they put a doll on a toy bed; they played with a car toy to generate actions like turning; and they stacked one toy on top of others, etc. While playing, parents also verbally described those manual actions generated by themselves or by infants.

Figure 1A: Experimental setup

Figure 1B: Examples from the infant egocentric view. The crosshair in each example indicates the infant's gaze direction.

Eye-tracker and Calibration

Parents and infants wore head-mounted eye trackers (Positive Science LLC). The tracking system has been successfully used in both infant and adult experiments (Franchak & Adolph, 2010; Yu & Smith, 2017). The eye-tracking system includes an infrared camera mounted on the head and pointed to the right eye of the participant that records eye images and a scene camera that captures and records images from the participant's perspective. The visual field of the scene camera is 108° (Figure 1B). Each tracking system – the infants' and parents' – recorded egocentric video and the x- and y-position of the right eye in the captured scene at a sampling rate of 30Hz. For eye-tracker setup, one experimenter engaged with the infant with an enticing toy while the second experimenter affixed the eye-tracker on the parent. After the parent's eye-tracker was secure and the scene and eye cameras were properly adjusted and oriented, both experimenters and the parent worked together to place the headgear and eye-tracker on the infant. The parent and one of the experimenters played with the infant while the other experimenter placed the infant's headgear (a small hat with Velcro stickers on the forehead) on the infant.

Instructions and Procedure

After the calibration phase, one of the experimenters distribute the set of toys on the floor and leave the parent and infant to play. The experimenters watch the interaction in an adjoining room and monitor the parent's and infant's eye and scene live streaming videos. If infants touch the camera or bumped the camera with a toy, the experimenter would go into the room, readjust the cameras, complete a new calibration phase, and leave the room so the parent and infant could complete the rest of the toy play session. Parents were asked to engage with their infants and toys as naturally as possible for ten minutes.

Data Annotation

Parent speech and infants' egocentric video were used in data analysis. We first transcribed speech and then identified spoken utterances containing action verbs. For those utterances, we further coded subject, verb, and (direct and/or indirect) object for each verb utterance. Since the main interest of this paper is on early verb learning and most verbs learned early by young children are action verbs, we focused only on action verbs with concrete meanings that can be revealed by manual actions (e.g. stack and shake) instead of abstract verbs (e.g. think and imagine).

For each parent utterance containing an action verb, we defined a window ranging from 3 seconds before to 3 seconds after the verb was generated. Within this temporal window, we first coded whether an action event was accompanied by the action verb, using infants' egocentric video. For example, in Figure 2, when a parent said, "shaking it", whether the parent or the infant used an object to generate a "shaking" action at the same time. If so, we next coded which target object was action-related for the action event. Figure 2 showed three example verb utterances, two accompanied by an action, and the other without any action. For the two utterances with action (Figure 2B, Row 1), we also coded target objects at the moment (Figure 2B, Row 2). Finally, an in-house coding program was used to code frame by frame which object infants attended moment by moment and gaze data were used to measure infants' attention when hearing verb utterances (Figure 2B, Row 3).

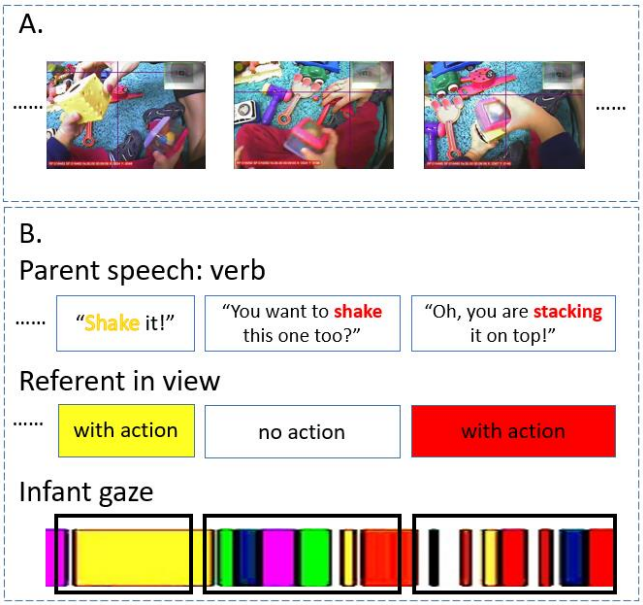


Figure 2: (A) Infants' first-person view and point of gaze during verb utterances. Purple crosshair in the image indicates where the infant was attending in the first-person view. (B) Row 1: Speech transcription of parents' verb utterances; Row 2: Coding of whether a verb utterance was accompanied by an action revealing the meaning of the verb. Colors indicate which object is carrying out the action; Row 3: Gaze coding. Different colors in the infant gaze stream indicate different objects attended by the infant moment by moment. If an infant attended to the named object, the colors in Row 2 and 3 would match in time.

Results

Verb utterance in parent speech. Parent speech contains 4406 utterances (1498 contain object names, 1381 contain action verbs). On average, parents generate roughly the same amount of nouns (5.09 nouns/min) and verbs (5.12 verbs/min, $t(32) = .72$, $p = .47$, ns). Among all the verbs, 705 were action verbs and 268 were abstract verbs. Thus, action verbs took roughly 72.4% of the total of 973 verbs, suggesting that parents most often used concrete verbs in their speech when they played with their children. Among all the action verbs that were coded from parent speech, we selected the top 25 verbs with relatively high frequency (except "look" and "see" as these two verbs were mostly used for attention getting in free play) to form a list of target action verbs for further data analysis. Figure 3A shows a skewed frequency distribution of those top 25 action verbs with two statistical properties. First, even for those top 25 verbs, most of them were produced fewer than 30 times, suggesting that a large proportion of those action verbs were hardly repeated by parents in a play session. Second, the skew distribution also revealed that some verbs were mentioned in parent speech much less frequently than others. Both the frequency difference between action verbs and object nouns, and the variability within action verbs suggest that those quantitative discrepancies are one of the many reasons why (some) verbs might be harder to learn than nouns.

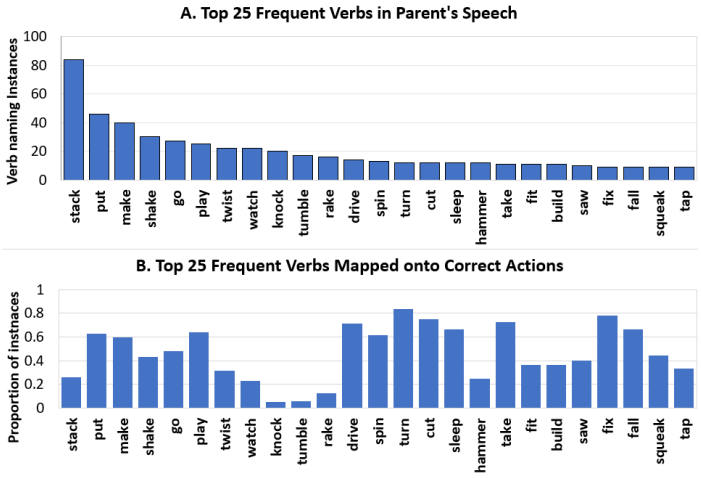
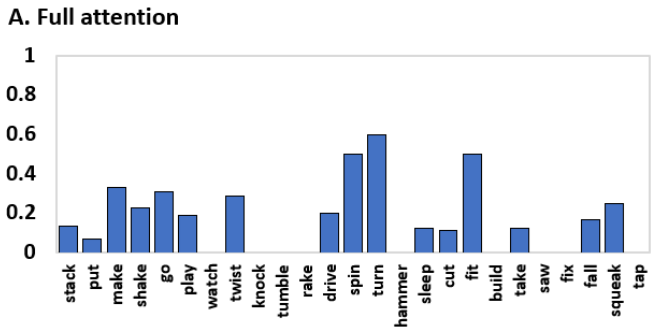


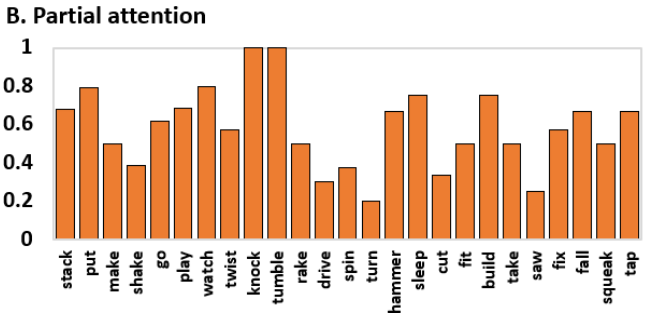
Figure 3: (A) A skewed frequency distribution of the top 25 action verbs. (B) A distribution of the percentage of verb utterances that were accompanied by an action.

Verb-Action co-occurrence. Learning the perceptually grounded meaning of an action verb requires not only hearing the verb but also perceiving the action. One critical question is how often a verb and its corresponding action co-occur in the learning environment? We answered this question by directly measuring verb-action co-occurrence and counting how often an action revealing the meaning of a verb was generated – either by parents or infants – when parents produced a verb utterance. Figure 3B showed the percentage of verb utterances that were accompanied by the corresponding action. There are two noticeable patterns. First, there is variability in verb-action co-occurrence across action verbs as when some verbs (e.g. "drive", "turn", "cut") were mentioned in parent speech, it was very likely that the corresponding actions were also generated at the same time; while for some other verbs (e.g. "knock", "tumble", "rake"), they were produced in parent speech most often without the corresponding actions. In those situations, either parents failed to demonstrate the corresponding action while a verb was generated, or parents failed to name the actions

conducted by infants. Second, there is no correlation between verb frequency and verb-action co-occurrence (r = -0.161, *n.s.*), suggesting that producing action verbs more frequently would not necessarily create more verb-action co-occurrences, which are critical for building verb-action mappings than just hearing action verbs alone.

Attention to verb-action co-occurrence. Infants' visual attention in free play was dynamic as they sometimes followed parents' attention and sometimes went with their own goals. Even with the presence of verb-action co-occurrence in the learning environment, they may or may not attend to the action when hearing a verb label. Given that attending the corresponding action when hearing a label is critical for verb learning, we next measured the proportion of infant gaze attention on the corresponding action within a verb utterance. Prior research shows that infants' learning of an object name depends on sustained visual attention to the object during a window that lasts from the onset of the utterance containing the name to several seconds after the offset of the utterance (Yu & Smith, 2012). Therefore, we operationally defined a verb event starting at the onset of a parent verb utterance and lasting for 3 seconds – the temporal interval including both the utterance itself (on average 1.5 sec long) and roughly 1.5 seconds after the utterance. We quantified infants' attention during and after hearing a verb utterance by defining three attentional states based on infant gaze: **Full attention** -- infants attended to the action 100% of time within a 3s window; **Partial attention** -- infants attended to the action sometimes but also to elsewhere when hearing a verb label. **No attention** -- infant did not attend to the action at all. Figure 4 showed the percentages of verb-action co-occurrences that received full attention (4A), partial attention (4B) or no attention (4C) from infants. As observed in the distributions of verb utterance frequency and verb-action co-occurrence, there is large variability among different action verbs. Infants seemed to attend to some actions (e.g. "turn" and "spin") much more than others (e.g. "saw" and "rake") when hearing verb labels. Also, in most cases, they seemed to attend to the correspond action sometimes but not the whole time within a 3s window as the percentages in partial attention are much higher than the percentages in full attention and no attention.





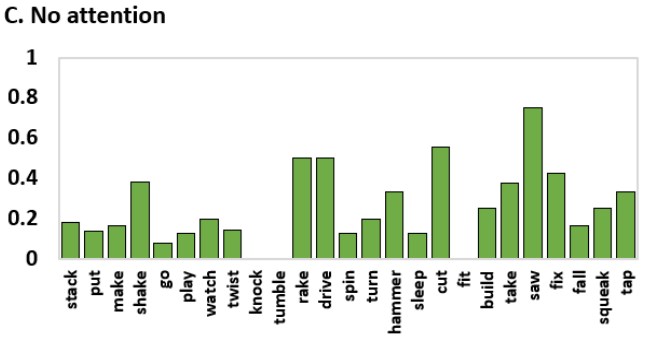


Figure 4: The distribution of infants' attention on the target of the actions that matches with the verbs of the top 25 verbs. (A): Infants' full attention on the targets; (B): Infants' partial attention on the targets; (C): Infants' no attention on the targets.

If an action verb was mentioned more frequently in parent speech, would a higher frequency attract infant attention more on the corresponding action when it was available in the environment? To answer this question, we correlated both verb utterance and verb-action co-occurrence with the three attentional states. As shown in Table 1, we found no correlation between verb frequency and infant attention. Producing more verb utterances did not attract infant attention more toward actions when those verb utterances were accompanied by the corresponding actions. However, there is a significant correlation between verb-action co-occurrence and full attention as shown in both Table 1 and Figure 5, suggesting that infants were more likely to attend to the action 100% of time when a verb and its corresponding action consistently co-occurred together. The higher percentage that manual actions and verb labels co-occurred together; the more likely infants showed full attention to the action event when hearing its label.

Table 1: The correlations between infants' attention, and verb utterance and verb-action co-occurrence (*p<0.05)

	verb utterance	verb-action co-occurrence
full attention	0.015	**0.475***
partial attention	0.206	-0.386
no attention	-0.242	-0.015

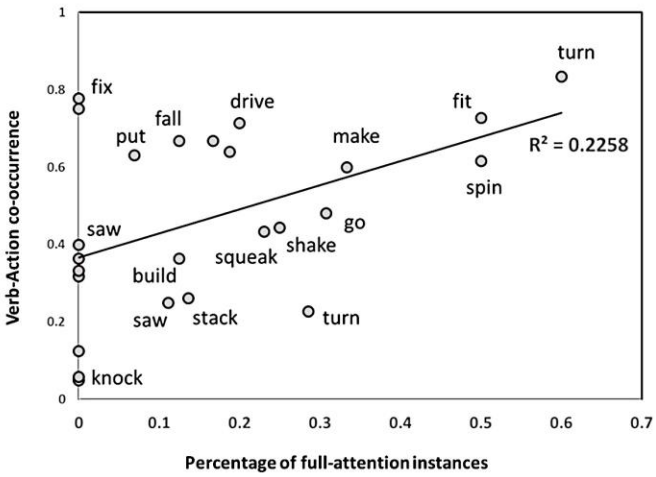


Figure 5: A significant correlation between verb-action co-occurrence and full attention.

Discussion

Recent studies show the overall speech input perceived by the young learner is predictive to later learning outcomes (Weisleder & Fernald, 2013). The input for learning concrete verbs includes not only spoken words but also their perceptually grounded meanings to build word-referent mappings. In light of this, the present study quantified both speech input and verb-action co-occurrences in the learning environment. Critically, the input to the language learning system is not the objective properties available in the learning environment but instead the information in the environment selected by the learner. Therefore, we measured the statistical regularities from the learner's perspective by using the learner's gaze data. Our result showed that the overall frequency distribution of verb generated by the parent during free play is right-skewed, which is similar to what has been observed in the recent studies of object names (Smith, Jayaraman, Clerkin, & Yu, 2018; Bambach, Crandall, Smith, & Yu, 2018). Moreover, the mere frequency of verb utterances was not related to how often the corresponding action was generated. It is not the case that more frequent verbs have more chances to be learned due to more frequent verb-action co-occurrences. In fact, some lower frequency verbs may have more chances to be learned as they co-occurred more frequently with the corresponding action. Further, it is not the case that more verb-action co-occurrences lead to more attention from the learner to the corresponding action.

The infant's attention adds a critical factor to variability of the learning input. We found that it is unlikely that young learners look at the target action during the entire time of a verb utterance. Instead, in most cases, they spent only some time looking at the corresponding action while hearing its label. It is also unlikely that they would completely miss the co-occurring action when a verb is heard. Although there isn't a significant correlation between verb frequency and infant attention, verb-action co-occurrence is positively correlated with the infant's full attention. For those verbs that co-occur more with the corresponding actions, infants are more likely to spend more time looking at the corresponding actions. The great variability within the concrete verbs examined here offers an explanation on why some concrete verbs are harder to learn than others.

What exactly makes verb learning difficult? Based on our findings, we argue that actions to which verbs refer are usually transient in context. Unlike concrete nouns whose perceptual information is usually available to the child when the object label is uttered, the corresponding action of a verb is not very likely to be perceptually available for the child continuously before, during, and after the verb utterance. Given the verb's transient nature and the infants' developing attentional system, if infants failed to attend to the right action at the time a verb was generated, they would miss the target action and once they miss the action, it is impossible for them to recover from other perceptual inspection of the immediate visual context at the moment. Despite the fact that verb learning is challenging, it is also important to keep in mind that verb learning happens in rich naturalistic contexts. Besides solely observing the action accompanying the verb, children also receive other cues that could help them figure out the correct mapping. For example, parents often provide socio-attentional cues, such as pointing to guide the child's attention (Goldin-Meadow, 2007). In addition, verbs are likely to co-occur with nouns and other parts of speech. Infants can also utilize the syntactic structure of the sentences to bootstrap the verb meaning (Naigles, 1996; Yuan, Fisher, & Snedeker, 2012).

The present study is the first step towards understanding the input for early verb learning. There are several future directions to advance our understanding on this topic. First, the current study does not have an outcome measure of verb learning as a way to directly assess the infant's knowledge of the heard verbs. Adding a verb learning test at the end of the play session would allow us to directly examine how the quantity and quality of co-occurrence statistics impact verb learning. Another way to link the input with learning outcomes is to collect and use the parent report of the child's vocabulary (i.e. MCDI, the MacArthur-Bates Communicative Development Inventories). Many studies have showed that both the quality and quantity of parent object naming are correlated with the child's MCDI results. However, little is known about how input quantity and quality impact early verb learning.

Second, toy play is only one of the everyday contexts in which children learn words. It would be interesting to study other learning contexts, such as storybook reading. Talking about objects on a page during book reading and manually manipulating objects during toy play are two very different types of interactions. Therefore, parent and children tend to generate very different learning statistics. Given that word-learning outcomes heavily depend on the structure of the input, it would be interesting to examine what types of input infants receive in those two contexts and compare how different types of input influence verb learning in those contexts.

Finally, another idea for follow-up studies is to compare the actions generated by infants versus by parents. There are studies showing that infants' own egocentric views contain unique properties and distributions that are critical for successful learning (Yurovsky, Smith, & Yu, 2013; Bambach, Crandall, Smith, & Yu, 2018). Actions generated by the parent may contain different visual properties from actions generated by the child. We could further investigate how the infant's body and associated visuomotor processes influence how the information is perceived and processed for learning verb-action mappings.

Acknowledgments

This research was supported in part by National Institutes of Health Grant R01HD074601 and R01HD093792 to CY and by the MOE (Ministry of Education in China) Project of Humanities and Social Sciences (Project No.16YJA190004), the National Natural Science Foundation of China (No. 71771102), and China Scholarship Council (CSC) (No. 201806775024) to SL. We would also like to thank Seth Foster, Anting Chen, Grace Lisandrelli, Lauren Slone, Drew Abney, and Daniel Pearcy, for data collection, and Emily Marie Heldman, Loren Louise Chastain, and Swasti Shree Singh, for data annotation.

References

Bambach, S., Crandall, D. J., Smith, L. B. & Yu, C. (2018). Toddler-Inspired Visual Object Learning. *Advances in Neural Information Processing Systems (NIPS), 31*.

Franchak, J. M. & Adolph, K. E. (2010). Visually guided navigation: Head-mounted eye-tracking of natural locomotion in children and adults. *Vision research, 50*(24), 2766-2774.

Gentner, D. (1982). Why nouns are learned before verbs: Linguistical relativity versus natural partitioning. In S. A. Kuczaj II (Ed.), *Language Development,* vol. 2: *Language, Thought, and Culture* (pp.301-334). Hillsdale, New Jersy: Lawrence Erlbaum.

Gentner, D. (2006). Why verbs are hard to learn. In K. Hirsh-Pasek & R. Golinkoff (Eds.), *Action Meets Word: How Children Learn Verbs,* (pp. 544–564). Oxford: Oxford University Press.

Gleitman, L. R. & Trueswell, J. C. (2018). Easy words: reference resolution in a malevolent referent world. *Topics in Cognitive Science*, 1-26.

Golinkoff, R.M., Chung, H. L., Hirsh-Pasek, K., Liu, J., Bertenthal, B., ……, Hennon, E. (2002). Young children can extend motion verb labels to point-light displays. *Developmental Psychology, 38*, 604–614.

Golinkoff, R. M., Hirsh-Pasek, K., Bailey, L., & Wenger, N. (1992). Young children and adults use lexical principles to learn new nouns. *Developmental Psychology, 28*, 99-108.

Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young American children*. Paul H Brookes Publishing.

Hirsh-Pasek, K. & Golinkoff, R. M. (2006). *Action Meets Verb: How Children Learn Verbs*. New York: Oxford University Press.

Hoff, E. (2003). The specificity of environmental influence: Socioeconomic status affects early vocabulary development via maternal speech. *Child development*, *74*(5), 1368-1378.

Imai, M., Li, L., Haryu, E., Okada, H., Hirsh-Pasek, K., Golinkoff, R. M., & Shigematsu, J. (2008). Novel noun and verb learning in Chinese-, English-, and Japanese-speaking Children. *Child Development, 79*(4): 979-1000.

Maguire, M. J., Hirsh-Pasek, K., Golinkoff, R. M., & Brandone, A. C. (2008). Focusing on the relation: fewer exemplars facilitate children's initial verb learning and extension. *Developmental Science, 11* (4): 628–634.

Messenger, K., Yuan, S., & Fisher, C. (2015). Learning verb syntax via listening: New evidence from 22-month-olds. *Language Learning and Development, 11*(4): 356-368.

Monaghan, P., Mattock, K., Davies, R. A., & Smith, A. C. (2015). Gavagai is as gavagai does: Learning nouns and verbs from Cross-Situational statistics. *Cognitive science*, *39*(5), 1099-1112.

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

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

Pulverman, R., Golinkoff, R. M., Hirsh-Pasek, K., & Buresh, J. S. (2008). Infants discriminate manners and paths in nonlinguistic dynamic events. *Cognition, 108*(3): 825-830.

Quine, W. V. O. (1960). Word and object (Studies in Communication). *New York and London: Tech-nology Press of MIT*.

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, L. B., Jayaraman, S., Clerkin, E. & Yu, C. (2018). The Developing Infant Creates a Curriculum for Statistical Learning. *Trends in Cognitive Sciences, 4*, 325-336.

Weisleder, A., & Fernald, A. (2013). Talking to children matters: Early language experience strengthens processing and builds vocabulary. *Psychological science*, *24*(11), 2143-2152.

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.

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

Yu, C. & Smith, L. B. (2017). Hand-eye coordination predicts joint attention. *Child Development, 88*(6): 2060-2078.

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