Learning semantic knowledge based on infant real-time attention and parent in-situ speech

Jane Yang (jane.yang@austin.utexas.edu)
Department of Psychology, University of Texas at Austin

Yayun Zhang (yayun.zhang@mpi.nl)
Language Development Department, Max Planck Institute for Psycholinguistics

Chen Yu (chen.yu@austin.utexas.edu)
Department of Psychology, University of Texas at Austin

Abstract

Early word learning involves mapping individual words to their meanings and building organized semantic representations among words. Previous corpus-based studies (e.g., using text from websites, newspapers, child-directed speech corpora) demonstrated that linguistic information such as word co-occurrence alone is sufficient to build semantically organized word knowledge. The present study explored two new research directions to advance understanding of how infants acquire semantically organized word knowledge. First, infants in the real world hear words surrounded by contextual information. Going beyond inferring semantic knowledge merely from language input, we examined the role of extra-linguistic contextual information in learning semantic knowledge. Second, previous research relies on large amounts of linguistic data to demonstrate in-principle learning, which is unrealistic compared with the input children receive. Here, we showed that incorporating extra-linguistic information provides an efficient mechanism through which semantic knowledge can be acquired with a small amount of data infants perceive in everyday learning contexts, such as toy play.

Keywords: language learning; semantic development; multimodal learning; egocentric vision

Introduction

Extensive research has explored how children acquire word knowledge. The majority of studies operate under the assumption that the primary objective in word learning is to establish a connection between a word and its correct referent or category of referents. While this mapping is undeniably crucial, the introduction of a new word also brings along a wealth of additional information. For example, a child encountering a toy helmet for the first time may hear:

Mother: Look! It's a helmet!

Mother: Football players wear helmets.

Mother: Helmets protect their heads.

The child needs to learn the association between the label helmet to its correct referent, but is also receiving information about related concepts (i.e., football players, head, protect, wear). Previous work has shown that toddlers have an early understanding of semantic relations (Arias-Trejo & Plunkett, 2013) and that their early environment is highly organized and capable of supporting such learning (Savic, Unger, & Sloutsky, 2023; Unger, Yim, Savic, Dennis, & Sloutsky, 2023; Savic, Unger, & Sloutsky, 2022; Huebner & Willits, 2018). Because parents' speech is often related to what toddlers are seeing and doing in the moment, children's linguistic input is semantically related and multisensory (Suanda, Smith, & Yu, 2016; Frank, Tenenbaum, & Fernald, 2013; Gogate, Bahrick, & Watson, 2000). Despite decades of work on semantic knowledge, it is still not clear how toddlers build semantic knowledge using their everyday linguistic and extra-linguistic input. The goal of this project is to understand how extra-linguistic information, such as infants' visual experience of seeing the same object across different learning moments can be used to group semantically related words in parent speech, and as a result, facilitate the learning of semantic relations.

Children as young as 24 months show an understanding of word relations. Many semantic priming studies have shown that children's processing of a spoken word may be delayed when presented with another object semantically related to it (Bergelson & Swingley, 2013), or it may be facilitated when the spoken word is preceded by another related spoken word (Arias-Trejo & Plunkett, 2013). In addition, auditory priming studies also provide evidence suggesting that young children can activate lexical semantic knowledge in the absence of visual referents or sentence contexts (Willits, Wojcik, Seidenberg, & Saffran, 2013). These studies provide strong and converging evidence that by the age of 2, children can successfully represent the semantic relations between words when processing language.

Both computational and experimental studies have found substantial semantic structures exist in natural speech and that human learners are sensitive to this information. Recent models applied to datasets with child-directed speech corpora, such as CHILDES (Huebner & Willits, 2018) show that linguistic information alone is sufficient to build semantically organized word knowledge, such as hierarchically structured categories (e.g., dog and beagle), associative links between words that co-occur frequently (e.g., blue and sky), and taxonomic links between words that belong to the same category (e.g., cat and dog). Importantly, these complex and highly organized semantic structures emerge automatically from the statistical regularities in child-directed speech without explicit training. This implies that the linguistic input children receive is highly structured, supporting the learning of linguistic knowledge at multiple linguistic levels. However, children's naturalistic language learning environment is

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

highly dynamic and multimodal. Infants in the real world hear words surrounded by contextual information, going beyond inferring semantic knowledge merely from language input. The extra-linguistic contextual information that children receive is likely to serve as an effective mechanism for acquiring semantic knowledge.

In the current study, we propose a computational mechanism through which infant real-time attention on visual objects can be used to group semantically related words in parent speech across multiple spoken utterances over time. Using parent speech mentioned at the beginning as an example, in that particular context, the child may see a toy helmet and build semantic relations between heard words like "helmet", "football player", "protect", "wear", and "head". In another context, the child may see the helmet again, but this time paired with another set of relevant "bag" of words. For instance, parents may describe the helmet's color, shape, and texture this time by introducing new words like "red", "round", "hard", etc. So each time the child sees the same object, the in-situ words provided by parents all contribute to their existing semantic network. This integration of infant attention and parent in-situ speech can provide children with additional visually grounded information to link related "bags" of words, which can be an efficient way for children to acquire semantic knowledge in everyday contexts.

To empirically test this idea, we recruited 15-to-24-month-old children and their parents and fitted them with head-mounted eye trackers to capture both the children's and the parents' first-person views while parents play toys with their children. This method allows us to track children's moment-by-moment attention focus, and identify sustained attention (SA) moments where the child gazed towards the same object for a duration of longer than three 3s. SA moments are important information-processing moments and have been found to be predictive of language learning (Yu, Suanda, & Smith, 2019). In the current study, we look at semantic networks built around children's SA episodes as a way to examine whether repeatedly seeing the same object while hearing different but related words facilitates the building of semantic networks among heard words.

In addition to SA, other modalities such as object handling (Tomasello & Farrar, 1986; Yu & Smith, 2012), also play important roles in organizing infant visual attention to objects (Deak, Krasno, Triesch, Lewis, & Sepeta, 2014; Yu & Smith, 2013, 2017). Studies have shown that manual exploration(especially from the child) can create visual saliency leading to increased visual attention on target, contributing to real-time information processing (Yang, Smith, Crandall, & Yu, 2023; Bambach, Smith, Crandall, & Yu, 2016). Furthermore, infants' bodily actions not only create visual data with unique properties in their first-person view but also elicit child-directed speech from responsive caregivers (Suarez-Rivera, Linn, & Tamis-LeMonda, 2022). As a result, infants learn the names of objects in their hands in both the laboratory and home environments (Suarez-Rivera et al., 2022; Yu & Smith, 2012), providing growing evidence that motor development is closely tied to early word learning (Yu & Smith, 2012; Iverson, 2010). Therefore, in our study, we also added manual actions from the child as an additional grounding modality and examined whether this type of additional contextual information contributes to semantic knowledge.

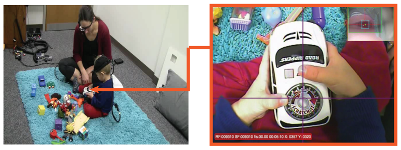
Figure 1: An overview of the experimental setup from third-person (left) and the infant's egocentric (right) views. Infants and their parents played freely with a set of toys. Both wore a head-mounted eye tracker which recorded gaze data. The purple cross-hair in the egocentric image indicates the infant's gaze point.

By linking extra-linguistic information such as children's real-time visual attention as well as active object manipulation, with the associated in-situ speech input, we aim to answer three research questions:

1. During multiple SA moments where children consistently attend to the same object, are the words children hear across multiple moments semantically related to each other?

2. During multiple SA moments where children consistently attend to the same object, are the words children hear across multiple moments semantically related to the label of the attend object?

3. Does having manual actions as an additional grounding modality create a stronger or weaker semantic network than having SA moments solely?

We hypothesize that incorporating extra-linguistic information provides an efficient mechanism through which semantic knowledge can be acquired with a small amount of data infants perceive in everyday learning contexts, such as toy play.

Method

Data collection

Participants The data used in this analysis were collected from 26 parent-child dyads who resided in midwest United States. All children were between the age of 15 to 24 months (M = 19.3, SD = 2.1, *Min* = 15.2, *Max* = 24.3) and were monolingual English speakers.

Materials Twenty-four everyday toys were selected for the experiment. Based on normative data, their names were expected not to be in the vocabulary of the infants but to be known to parents. Parents were allowed to freely refer to the toys by any or multiple different labels. Toys were selected so that multiple unique objects belonged to the same category.

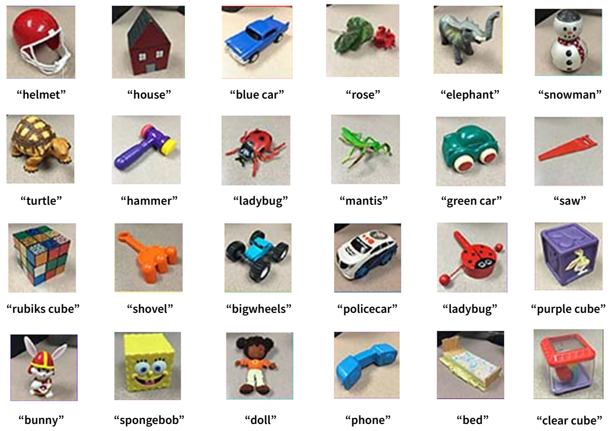

Figure 2: This plot displays twenty-four toy objects and the most frequent object name used by parents.

For example, the "vehicle" category included toys "car", "police car", and "truck"; the "insect" category included toys "ladybug" and "mantis". Details of all twenty-four toys used in the experiment can be found in Figure 2.

Experimental setup During a play session, the parent and infant sat next to each other on a carpet and were provided with 24 to play with. They each wore a head-mounted eye-tracker with a front-facing camera capturing their egocentric view. An example of the infant's view can be seen in Figure 1-right (Franchak, Kretch, Soska, Babcock, & Adolph, 2010; Yu & Smith, 2013). The eye camera was mounted on the head and pointed to the right eye of the participant. The scene camera captured the first-person view from the participant's perspective, with a 90° horizontal field. Each eye tracking system recorded both the egocentric view video and gaze direction in that view, with a sampling rate of 30 Hz. Three third-person view cameras were also used to record the play session from a distance.

Procedure Before the toy play session, the experimenter randomly spread the 24 toys on the floor. When the family is ready to begin, the experimenter first fits the parent with the eye-tracking gear. The experimenter then gives the dyad some toys and asks the parent to engage the child with toys while experimenter one puts the eye-tracking gear low on the forehead of the child and adjusts its position.

After both the parent's and the child's eye-tracking gears are placed properly, Two experimenters collect calibration points for eye-tracking. Experimenter one randomly points to toy objects in the play area using a laser pointer and experimenter two makes sure both the parent and the child's attention is directed to that point. This procedure is repeated at least 15 times so we have enough calibration points collected for offline calibration later.

Parents are then instructed to play with the provided toys as they would at home and to keep their children engaged with those toys. Together, each session in our final dataset lasted an average of 7.01 minutes (range 3.74 - 11.69 min), with 350,042 image frames each extracted from the infant's and the parent's egocentric views (30 frames per second).

Data processing

Synchronization and calibration Egocentric videos, eye videos, and third-person view videos were first synchronized in time and decomposed into image frames. We then followed a calibration procedure commonly used in head-mounted eye tracking (the details provided in Yu, Zhang, Slone, and Smith, 2021). After calibration, a cross-hair was superimposed in each of the egocentric images to indicate the wearer's visual attention in view. An example of a calibrated egocentric view with crosshair is shown in Figure 1-left. In total, 78.4% of frames from infant's view contain ROIs to toy objects (274,539 frames of infant's gaze data in total).

Data annotation From calibrated videos, we annotated three types of behaviors:

> **Gaze direction**: Each of the 24 toys in toy play sessions was identified as a region-of-interest (ROI). Coders watched the calibrated egocentric videos frame-by-frame and coded an ROI for each of the frames using an in-house program (Figure 3, row 1).
>
> **Manual holding**: Coders watched a play session from the views of multiple cameras and annotated, frame-by-frame, the object with which the infant's hands made contact. Coders went through the session twice, once to annotate manual action from the left hand, and then to annotate the right hand. In total, there were 1,234 instances of infants' manual action events (*M* = 47.46, *SD* = 30.15).
>
> **Parent speech**: We used WhisperX, a speech recognition algorithm to automatically transcribe parents' speech into spoken utterances (Bain, Huh, Han, & Zisserman, 2023). As shown in the third row in Figure 3, each utterance is defined as a string of speech between two periods of silence lasting at least 400ms (Yu & Smith, 2012).

Measures

Sustained attention We defined sustained attention as the stabilization of visual attention to the same object for durations longer than 3 seconds. We identified infant sustained attention events by measuring the frames the infants gazing towards the same object. Frames where infants attended to target objects were coded by human coders. We then filtered out attention bouts shorter than 3 seconds to obtain sustained attention bouts (see Figure 3 row 1 and 2).

"Bag" of words around sustained attention We temporally aligned the speech utterances with infant sustained attention so that utterances, when infants attended to the same toy, were grouped into the same "bag" of words. As shown in the second and third row in Figure 3, in cases where the child is attending to the helmet across three SA moments, the child hears the parent saying "here is a helmet", "football players wear helmet", and "helmet protects their head". Three speech utterances each aligns with the SA moments are

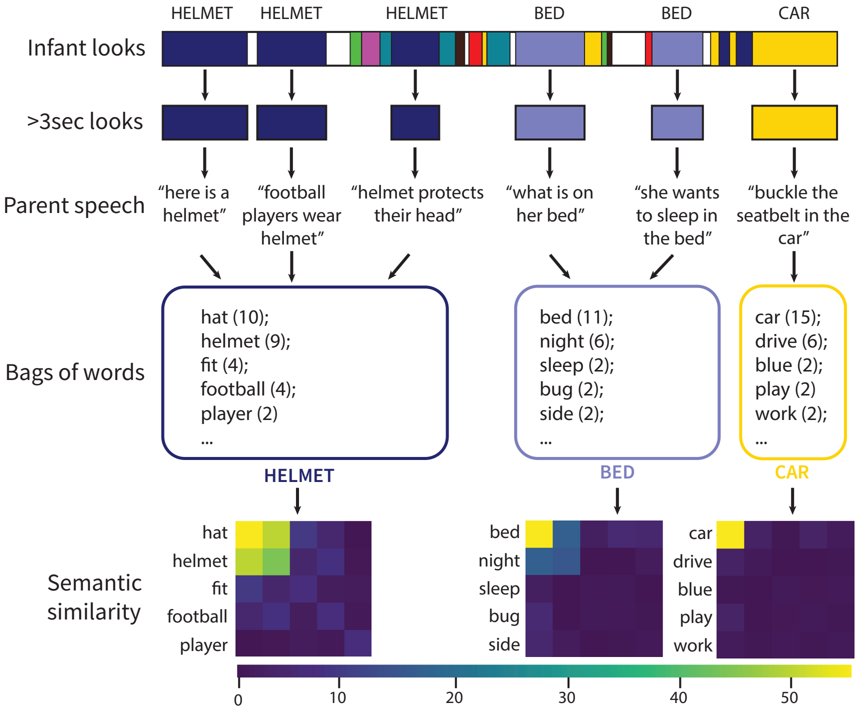
HELMET
HELMET
HELMET
BED
BED
CAR
Infant looks
>3sec looks
Parent speech
"here is a helmet"
"football players wear helmet"
"helmet protects their head"
"what is on her bed"
"she wants to sleep in the bed"
"buckle the seatbelt in the car"
Bags of words
hat (10); helmet (9); fit (4); football (4); player (2) ...
bed (11); night (6); sleep (2); bug (2); side (2); ...
car (15); drive (6); blue (2); play (2) work (2); ...
HELMET
BED
CAR
Semantic similarity
hat
helmet
fit
football
player
bed
night
sleep
bug
side
car
drive
blue
play
work
0
10
20
30
40
50

Figure 3: An overview of the method used in the present study, demonstrating the formation of "bags" of words and the data analysis pipeline. The top temporal data stream represents the raw regions-of-interest (ROIs) data from infant gaze in toy play, where each color in the data stream represents a unique ROI. We first identified the infant sustained attention (defined as an unbroken look at the same object longer than 3s) shown in the second data stream. We then grouped utterances that were temporally aligned with them into each "bag" of words and calculated their frequency, shown in the fourth row. Lastly, the heatmaps at the bottom show the top five words' pairwise similarities.

grouped into the same "bag" of words (see the fourth row in Figure 3). We created a total number of twenty-four "bags" of words grouped by the twenty-four unique toys provided in the experiment. Each "bag" of words was further cleaned by removing function words without semantic meanings (e.g., "this", "that", "a", "the"; see Figure 3, row 4).

Shuffled baseline A baseline condition was created by temporally shuffling the infant sustained attention data and then grouping temporally aligned utterances based on shuffled attention. The total duration of data was preserved in the shuffled condition for a fair comparison.

Results

In the following analyses, (1) we will first address how children learn about the semantic relations between words that occurred across multiple SA moments on the same object. To do so, we will calculate the pairwise semantic similarity of visually grounded words infants hear during SA moment; (2) We then will address the question of how infants learn the mapping of individual words to their referents by calculating the individual word semantic similarity to the attended object; (3) Lastly, we will investigate the role of manual actions as an additional grounding modality for learning semantic knowledge, by comparing semantic similarities derived from SA bouts with or without manual actions.

Semantic similarity between words children hear during SA moments

To understand the role of infant sustained attention in learning semantic knowledge, we examined the contextual information provided by infant sustained attention. To do this, we first identified all possible word pairs that words can form in each "bag" of words and quantified each word pair's frequency.

To measure semantic similarity, we calculated pairwise similarities of all words in the same "bag" weighted by each pair's frequency. We found that words in the same "bag" grouped by infant attention are significantly semantically closer to each other than those grouped in shuffled condition ($M_{original}$ = 0.42, $M_{shuffled}$ = 0.24, t = 5.11, $p < 0.001$, see Figure 4A).

This result suggests that infants are likely to learn the association between visually grounded words that occur across different SA moments. When infants look at an object, they are more likely to hear words that are associated with each other. For example, if an infant looks at a toy helmet and hears the parent saying: "football players wear a helmet to protect their head", the infant is more likely to learn the associative among word pairs: "football" and "player", or "protect" and "head".

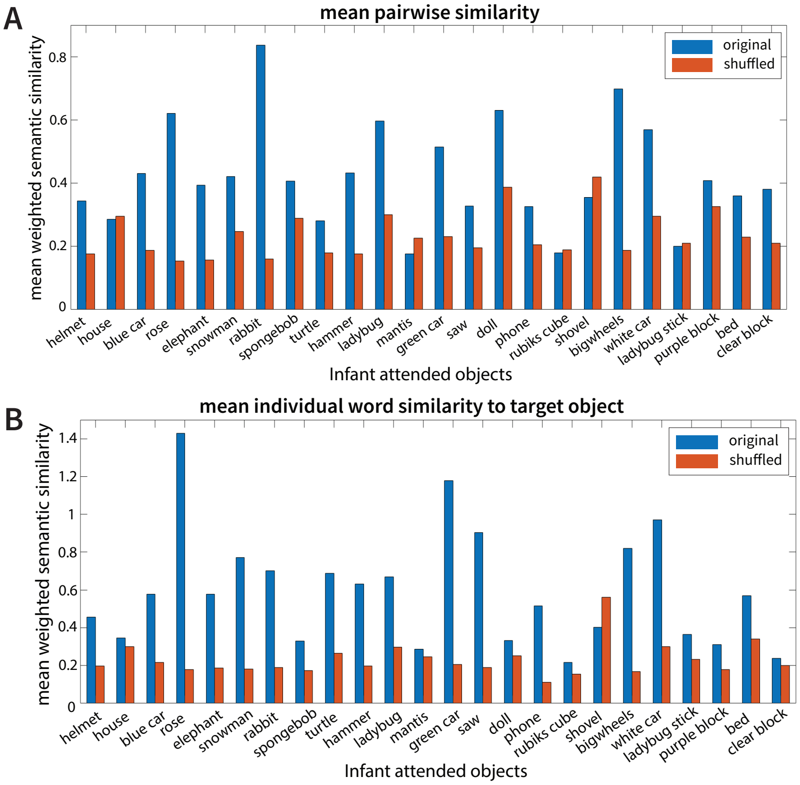

Figure 4: Plot A displays the mean pairwise similarities of all the words in each word "bag" weighted by each pair's frequency. Plot B displays the mean similarities between individual words in the "bag" to the target object of the same bag weighted by words' frequency. The blue bars represent the original group, and the orange bars represent the shuffled group.

Semantic similarity between words children hear and objects they attend

Next, we investigated whether grouping words based on sustained attention is an effective learning mechanism for infants to learn words that are directly related to the attended object. We first calculated the frequency distributions of individual words in each "bag", in which each word frequency would be used as a weight in semantic similarity calculation later. We then defined one naming label for each object referent. For referents that had multiple possible naming labels, the label with the highest frequency was used. For example, naming labels such as "truck", "vehicle", or "bigwheels" could all be possible labels for the object "truck"; however, the label "truck" was chosen given it was the most frequently mentioned by parents.

We then calculated the individual word's semantic similarity to the label of the attended object for each word in the same "bag", weighted by individual word frequency. We found that words grouped based on the same attended object are more semantically related to that object name compared with words in shuffled condition ($M_{original}$ = 0.60, $M_{shuffled}$ = 0.23, t = 5.62, $p < 0.001$, see Figure 4B).

This result suggests that connecting words based on visual grounding provides an effective computational approach to implementing how a young learner may link different word labels heard in similar contexts.

Comparing semantic similarities between SA moments with vs without manual actions

When an infant holds and looks at an object, what we refer to as hand-eye coordination, they create a stable, centered visual field, effectively reducing the visual clutter of competing objects (Bambach, Crandall, & Yu, 2013; Yu & Smith, 2012), providing optimal moments for learning to occur (Pereira, Smith, & Yu, 2014). We derived hand-eye coordination events by measuring the frames that included the infant holding on to and gazing towards the same toy. To examine whether manual actions during sustained attention events provide an additional grounding modality that can help infants learn semantic knowledge, we performed a side-by-side comparison of semantic similarities derived from two different types of SA moments.

We divided all sustained attention (SA) bouts into two mutually exclusive subsets: SA bouts with manual actions and SA bouts without manual actions (see Figure 5). There were comparable SA bouts with manual actions than without manual actions ($M_{SA\_with\_manual\_action}$ = 23.42, $M_{SA\_without\_manual\_action}$ = 23.54, t = -0.03, *ns*). Using the same approach described in the previous section, we temporally aligned speech utterances with SA with manual action bouts so that utterances, when infants held onto and looked at the same toy, were grouped into the same "bag" of words. Comparable numbers of aligned utterances were found for SA bouts with or without manual action ($M_{SA\_with\_manual\_action}$

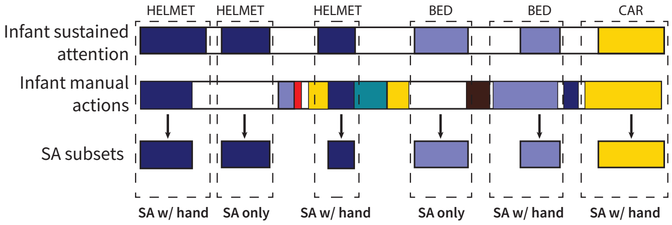

Figure 5: An overview showing how we divide infant sustained attention bouts into two mutually exclusive subsets: SA bouts with or without manual actions.

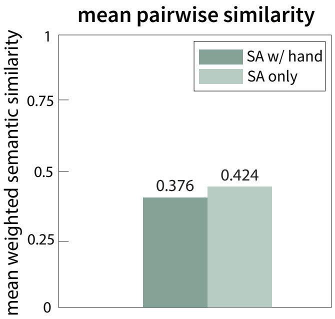

Figure 6: This shows the side-by-side comparison of pairwise word semantic similarity of all words in each "bag" of words grouped by SA bouts with or without manual actions.

= 37.29, $M_{SA\_without\_manual\_action}$ = 40.04, t = -0.18, *ns*).

We performed a parallel semantic similarity calculation on each "bag" of words grouped by sustained attention with manual action events. We first calculated the pairwise word semantic similarity of all words in the same "bag", weighted by each pair's frequency. We found that words grouped by either SA bouts with or without manual action were equally semantically close to words in the same "bag" ($M_{SA\_with\_manual\_action} = 0.376$, $M_{SA\_without\_manual\_action} = 0.424$, t = -1.09, *ns*, see Figure 6). We then calculated the individual word semantic similarity to the held and attended object, weighted by individual word frequency. Similarly, We found that words grouped by either SA bouts with or without manual action were equally semantically close to the target object ($M_{SA\_with\_manual\_action} = 0.591$, $M_{SA\_without\_manual\_action} = 0.595$, t = -0.06, *ns*, see Figure 7).

Taken together, the results suggest that having manual action as an additional grounding modality does not supplement more extra-linguistic information that can potentially allow infants to acquire semantic knowledge more effectively, compared to learning semantic knowledge purely from words that are visually grounded.

Discussion

The present study found that words in the same "bag" grouped by infant attention are semantically closer to each other as well as to the label of the attended target compared to those grouped in shuffled conditions. This result suggests that visual grounding of attention may provide useful cues to help children link the related heard words more easily, and as a result, facilitate their learning of semantic relations (e.g., the association between "football" and "player" when looking at "helmet"). Incorporating manual actions as an additional language grounding modality does not provide a more effective way for infants to learn about semantic knowledge.

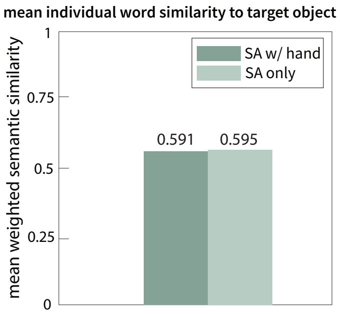

Figure 7: This shows the side-by-side comparison of individual word semantic similarity to the target object in each "bag" of words grouped by SA bouts with or without manual actions.

This study reveals a computational mechanism that utilizes infants' real-time attention to visual objects to cluster semantically related words in parental speech across multiple learning moments where the child attends to the same object. The results suggest that infant attention and parental spontaneous speech serve as effective means for infants to acquire semantic knowledge in their daily environments.

In the future, we plan to establish whether the time-locked linguistic structure identified in learning input can be used to learn object names and build semantic networks. Can natural scenes, early visual experiences with the objects, and a mix of more and less closely related bags of words lead to different learning outcomes? We will design a series of controlled experiments using the well-established cross-situational learning paradigm (Smith & Yu, 2008; Vlach & Johnson, 2013; Suanda, Mugwanya, & Namy, 2014; Fitneva & Christiansen, 2017). There is now substantial evidence showing that learners can accumulate information across multiple learning situations to learn word-referent mappings. However, it is still not clear how learners keep track of multiple to-be-learned words and multiple potential referents and how they build semantic networks within and across-situationally in naturalistic everyday situations like toy play.

In addressing this question, we will also examine the contributions of consistency versus diversity in the training set. In general, contextual diversity and interleaving of training instances increase learning and retention, and this has been shown in some cross-situational learning experiments as well. But both theory and evidence suggest that for novices and early stages of learning, consistency may be more important.

Acknowledgments

This work was supported by NICHD R01HD104624 and R01HD093792. We thank the Developmental Intelligence Lab for the feedback on earlier versions of this work.

References

Arias-Trejo, N., & Plunkett, K. (2013). What's in a link: Associative and taxonomic priming effects in the infant lexicon. *Cognition*, *128*(2), 214–227.

Bain, M., Huh, J., Han, T., & Zisserman, A. (2023). Whisperx: Time-accurate speech transcription of long-form audio. *INTERSPEECH 2023*.

Bambach, S., Crandall, D. J., & Yu, C. (2013). Understanding embodied visual attention in child-parent interaction. In *2013 ieee third joint international conference on development and learning and epigenetic robotics (icdl)* (pp. 1–6).

Bambach, S., Smith, L. B., Crandall, D. J., & Yu, C. (2016). Objects in the center: How the infant's body constrains infant scenes. In *2016 joint ieee international conference on development and learning and epigenetic robotics (icdl-epirob)* (pp. 132–137).

Bergelson, E., & Swingley, D. (2013). Young toddlers' word comprehension is flexible and efficient. *PloS one*, *8*(8), e73359.

Deak, G. O., Krasno, A. M., Triesch, J., Lewis, J., & Sepeta, L. (2014). Watch the hands: Infants can learn to follow gaze by seeing adults manipulate objects. *Developmental science*, *17*(2), 270–281.

Fitneva, S. A., & Christiansen, M. H. (2017). Developmental changes in cross-situational word learning: The inverse effect of initial accuracy. *Cognitive Science*, *41*, 141–161.

Franchak, J. M., Kretch, K. S., Soska, K. C., Babcock, J. S., & Adolph, K. E. (2010). Head-mounted eye-tracking of infants' natural interactions: a new method. In *Proceedings of the 2010 symposium on eye-tracking research & applications* (pp. 21–27).

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.

Gogate, L. J., Bahrick, L. E., & Watson, J. D. (2000). A study of multimodal motherese: The role of temporal synchrony between verbal labels and gestures. *Child development*, *71*(4), 878–894.

Huebner, P. A., & Willits, J. A. (2018). Structured semantic knowledge can emerge automatically from predicting word sequences in child-directed speech. *Frontiers in Psychology*, *9*, 133.

Iverson, J. M. (2010). Developing language in a developing body: The relationship between motor development and language development. *Journal of child language*, *37*(2), 229–261.

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

Savic, O., Unger, L., & Sloutsky, V. M. (2022). Exposure to co-occurrence regularities in language drives semantic integration of new words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *48*(7), 1064.

Savic, O., Unger, L., & Sloutsky, V. M. (2023). Experience and maturation: The contribution of co-occurrence regularities in language to the development of semantic organization. *Child development*, *94*(1), 142–158.

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

Suanda, S. H., Mugwanya, N., & Namy, L. L. (2014). Cross-situational statistical word learning in young children. *Journal of experimental child psychology*, *126*, 395–411.

Suanda, S. H., Smith, L. B., & Yu, C. (2016). More than words: The many ways extended discourse facilitates word learning. In *Cogsci*.

Suarez-Rivera, C., Linn, E., & Tamis-LeMonda, C. S. (2022). From play to language: Infants' actions on objects cascade to word learning. *Language Learning*, *72*(4), 1092–1127.

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

Unger, L., Yim, H., Savic, O., Dennis, S., & Sloutsky, V. M. (2023). No frills: Simple regularities in language can go a long way in the development of word knowledge. *Developmental Science*, *26*(4), e13373.

Vlach, H. A., & Johnson, S. P. (2013). Memory constraints on infants' cross-situational statistical learning. *Cognition*, *127*(3), 375–382.

Willits, J. A., Wojcik, E. H., Seidenberg, M. S., & Saffran, J. R. (2013). Toddlers activate lexical semantic knowledge in the absence of visual referents: Evidence from auditory priming. *Infancy*, *18*(6), 1053–1075.

Yang, J., Smith, L., Crandall, D., & Yu, C. (2023). Using manual actions to create visual saliency: an outside-in solution to sustained attention and joint attention. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 45).

Yu, C., & Smith, L. B. (2012). Embodied attention and word learning by toddlers. *Cognition*, *125*(2), 244–262.

Yu, C., & Smith, L. B. (2013). Joint attention without gaze following: Human infants and their parents coordinate visual attention to objects through eye-hand coordination. *PloS one*, *8*(11), e79659.

Yu, C., & Smith, L. B. (2017). Multiple sensory-motor pathways lead to coordinated visual attention. *Cognitive science*, *41*, 5–31.

Yu, C., Suanda, S. H., & Smith, L. B. (2019). Infant sustained attention but not joint attention to objects at 9 months predicts vocabulary at 12 and 15 months. *Developmental science*, *22*(1), e12735.