Using an Egocentric Human Simulation Paradigm to quantify referential and semantic ambiguity in early word learning

Spencer Caplan[1], **Misty Peng**[1], **Yayun Zhang**[2], **Chen Yu**[1]

spencer.caplan@austin.utexas.edu, misty.peng@utexas.edu
yayun.zhang@mpi.nl, chen.yu@austin.utexas.edu

[1]Department of Psychology, University of Texas at Austin, USA
[2]Language Development Department, Max Planck Institute for Psycholinguistics, NL

Abstract

In order to understand early word learning we need to better understand and quantify properties of the input that young children receive. We extended the human simulation paradigm (HSP) using egocentric videos taken from infant head-mounted cameras. The videos were further annotated with gaze information indicating in-the-moment visual attention from the infant. Our new HSP prompted participants for two types of responses, thus differentiating referential from semantic ambiguity in the learning input. Consistent with findings on visual attention in word learning, we find a strongly bimodal distribution over HSP accuracy. Even in this open-ended task, most videos only lead to a small handful of common responses. What's more, referential ambiguity was the key bottleneck to performance: participants can nearly always recover the exact word that was said if they identify the correct referent. Finally, analysis shows that adult learners relied on particular, multimodal behavioral cues to infer those target referents.

Keywords: word learning; human simulation paradigm; eye tracking; egocentric video; referential and semantic ambiguity

Introduction

How do infants learn their first words? A large literature tackles this problem from the perspective of conceptual development. Yet whether core conceptual knowledge precedes early word learning (e.g. Gleitman, Cassidy, Nappa, Papafragou, & Trueswell, 2005; Spelke & Kinzler, 2007) or is generated by this process (e.g. Waxman & Markow, 1995), the knowledge infants possess when growing their nascent vocabulary is radically limited compared with older children. This naturally constraints both *what* and *how* infants initially learn — and places heightened importance on understanding the actual perceptual input that infants receive, both its ambiguities and its regularities. Since infant word learning begins at just a few months of age (Bergelson & Swingley, 2012), and a large proportion of early vocabulary consists of concrete nouns, a key challenge in early word learning is to map heard names with seen objects.

Building word-object mappings in early word learning depends both on internal cognitive processes and the input on which those processes operate (Smith, Jayaraman, Clerkin, & Yu, 2018; Yang, 2004). In order to more fully understand learning in the real world, we need to precisely quantify properties of the input that feeds into the learning system (Bergelson et al., 2019); as they may differ substantially from controlled experiments (e.g. Xu & Tenenbaum, 2007; Yu & Smith, 2007). Young children, after all, learn words not in a laboratory but in everyday contexts. The learning mechanisms responsible for real-world learning may differ from perceptual and cognitive processes that operate on well-controlled stimuli (Dupoux, 2018). Some of the pioneering work on quantifying early input for word learning comes from Gillette, Gleitman, Gleitman, and Lederer (1999) and their Human Simulation Paradigm (HSP): in which adult (or child (Medina, Snedeker, Trueswell, & Gleitman, 2011)) participants are tasked with guessing the word uttered during a silenced video vignette extracted from video recordings of parent-child interaction. Gillette et al. (1999)'s primary finding is that naturally occurring environments are highly ambiguous: adult participants (with their fully-developed mental repertoire) struggle to correctly recover the original word for a majority of video snippets. Thus, young children's slowness in word learning prior to the vocabulary spurt (Clark, 1995) stems from the inherent ambiguity of mapping words to their labels in natural environments rather than infants' supposed lack of conceptual knowledge.

Locus of ambiguity

The fact that the perceptual input that children learn from is ambiguous (Gillette et al., 1999; Medina et al., 2011) is not in doubt. Yet this does not address the locus of ambiguity, nor explain how adult learners (let alone infants) extract information from that varied and multifaceted input. One obvious form of ambiguity relates to referents: if a child hears /wug/ when walking in the park, this may refer to the bird, or the squirrel, or the tree — who's to say! Yet learning is not finished even if the intended referent is correctly identified (Quine, 1960). Even if an additional cue like pointing clarifies that the pet in the room (and not the dishwasher) is the target of /dɔg/, the space of possible meanings for that phonological label could still be quite large — the word may be the particular pet's name, or it could mean pets generally. It might pick out the set of (all and only) dogs. But it also might select the set of poodles, or mammals, or animals. It might refer to the appearance of the dog: spotted or four-legged or tired. The list goes on. Beyond referential ambiguity, learners face an additional challenge of semantic ambiguity (Fodor, 1983). It is of course this intended conceptual relation that is the ultimate target of a word's meaning rather than simply a scene-dependent referent (Gleitman, 1990). While both types of ambiguity are certainly problems that a learner must solve, it

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society.*

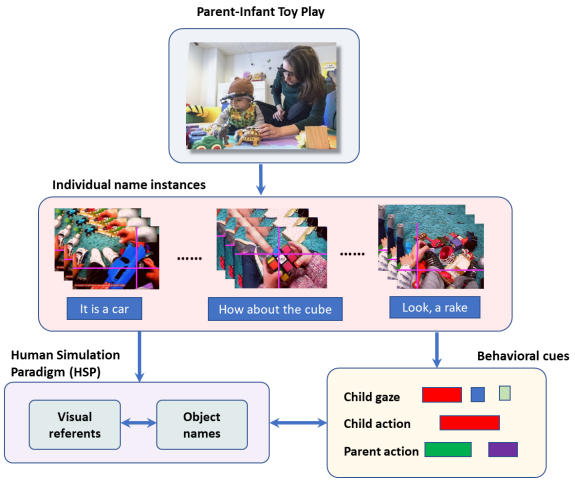


Figure 1: An overview of our egocentric Human Simulation Paradigm (HSP). Video vignettes are extracted from a toy-play corpus (Yu et al., 2021), and shown to adult learners to measure beliefs about potential referent and labels. Behavioral cues are annotated from those videos to analyze how those cues guide participants' guesses with respect to the target in each vignette.

is not clear which is the primary factor in driving/constraining children's earliest words, which do largely comprise concrete nouns (Gentner, 1982). In this paper, we aim to quantify properties of the input to address this question: is referential or semantic ambiguity the primary bottleneck for early word learning? To accomplish this, we extended the standard HSP by separately probing participants for referent and label responses after watching a vignette extracted from a naming moment of parent-child social interaction.

What I saw is what you get

The key rationale behind the HSP is to put adult learners into a similar learning situation as children experience; we can use adult learning performance to quantity the informativeness of the situation for infant learning. However, the original HSP does not perfectly serve the purpose because the videos used in the experiments are taken from a third-person perspective. Recent studies using head-mounted cameras show that such third-person views differ substantially from the egocentric views which children actually perceive (Kretch, Franchak, & Adolph, 2014; Smith, Yu, & Pereira, 2011). The only pertinent view for infant learning should come from the infant's own perspective. In fact, adult participants perform better in a cross-situational learning task when stimuli are taken from egocentric rather than third-person videos recorded in the same toy-play session (Yurovsky, Smith, & Yu, 2013; Zhang, Yurovsky, & Yu, 2021).

Unlike third-person videos, egocentric videos better approximate what original viewers actually perceived. However, participants watching egocentric videos in an HSP are free to examine any part of the video with high visual acu-

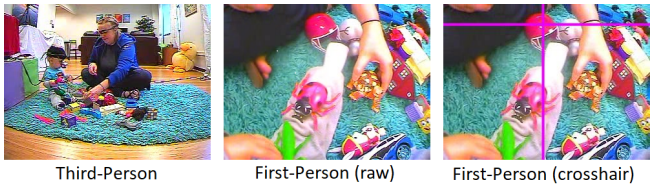


Figure 2: Examples of a third-person vs. first-person view (with and without a gaze crosshair imposed); these images were all taken at the same timestamp from a parent-infant dyad in (Yu et al., 2021) yet show radically different information for a potential learner.

ity. The original perceivers had their visual experience filtered through their attention system with a shockingly narrow band of central, foveal focus (Rosenholtz, 2016). To better approximate the perceptual input received by infant learners, our egocentric HSP not only makes use of egocentric videos recorded from parent-child free-flowing toy-play (Figure 1), but the infant's gaze direction is also provided in these vignettes via an explicit crosshair, providing a direct cue to their visual attention (Figure 2). This egocentric HSP is thus as close as possible to capturing the perceptual input that infants may receive during toy-play, a typical, everyday learning context. We also extracted from these vignettes both ground-truth information about the intended referent in parent speech as well as behavioral cues, such as the proportion of time that each target was being visually attended by the original infant or being held by either the infant or parent. This allows us to study how participants make use of (or do not make use of) different behavioral cues to resolve ambiguity in the input. Ultimately we are concerned not with which conditions best support learning, but understanding what in-the-moment behavioral cues participants (and infant learners) use to make their disambiguation choices. These analyses point to concrete directions for further experimental manipulation to causally assess the impact of such cues on learning.

Egocentric HSP Experiment

Our egocentric HSP improves on previous studies in two ways. First, the stimuli are first-person vignettes which contain a crosshair providing an explicit cue to the infant's original moment-by-moment visual attention. This paper is the first to include gaze information in the input for an HSP. Second, the experiment is further novel by prompting the participants to separately guess both the referent and the label from each video. This allows us to differentiate cases of referential ambiguity (selecting the right object from the scene) from semantic ambiguity (reconstructing the original conceptual characterization of that object).

In our egocentric HSP, participants see a series of videos extracted from an infant-parent toy-play corpus (Yu et al., 2021). Each extracted video is a five second clip centered on naming instances of twelve target words (*"bed," "SpongeBob," "snowman," "hammer," "turtle," "shovel," "phone,"*

"house," "helmet," "elephant," "doll," and *"rabbit"*). The audio from each video was removed and replaced by a beep at the moment that naming originally occurred (i.e. at the three second mid-point). In order to create a balanced sample from the full toy-play corpus (which contains 1,508 parent naming events), we extracted a quasi-random sample of between 12 and 15 videos for each of the twelve target words (owning to the different corpus-frequencies of each word). This resulted in a total of 172 stimulus vignettes overall. Our stimuli did not include cases where the same name was used repeatedly in an adjacent time window by parents; since while such instances are potentially distinct naming events, they do not represent unique contexts or unique perceptual input for participants.

Participants

Eighty-one participants were recruited from Amazon Mechanical Turk (MTurk) and paid $3.50 for their participation. Participants were divided between six semi-randomized blocks which varied the order and the exact stimuli presented. Each block contained 58 vignettes, and blocks were balanced to ensure we gathered the same number of judgements per video, while varying the trial-order within the experiment. Ten subjects who (due to a back-end MTurk issue) may have repeated the experiment more than once were excluded. Since this was done by logging and comparing participants' IP addresses, it is possible that some or all of these were unique participants who simply happened to share an IP address. Nevertheless they were excluded out of caution.[1] This resulted in seventy-one participants whose data was included in the final analyses.

Materials

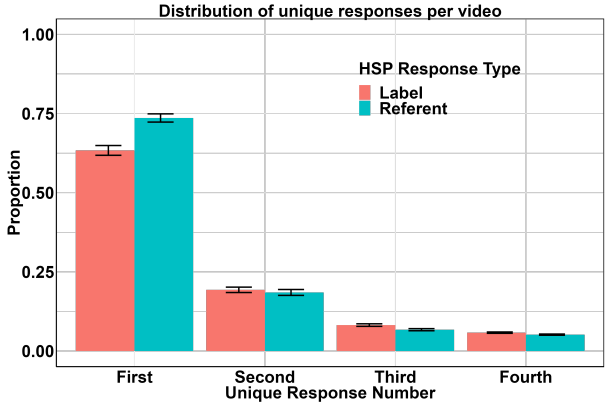
Stimuli were extracted from the video corpus presented in Yu et al. (2021). This naturalistic toy-play environment included 24 toys, varying in terms of properties such as size, shape, and color. Data from (Yu et al., 2021) included videos from 36 infant-parent dyads (mean infant age of 19.3 months) playing for 5-10 minutes (mean play time of 6.32 minutes.) The parents were not provided with particular instructions about how they should play with their infants, nor were they told to use particular names for any of the toys in the experiment room (indeed, they sometimes used more than one name to refer to the same toy.) The overall corpus consists of 1,508 spontaneous parent naming events, from which we extracted 172 stimulus vignettes as described above. Each vignette is a silenced, five second clip centered around the naming (indicated with a beep) of one of twelve target words.

Procedure

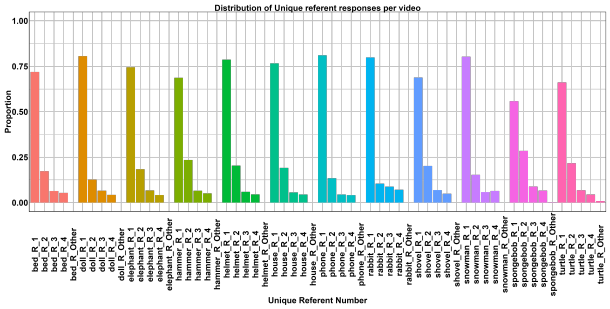
Participants completed the experiment via their web browser. The experiment was written using custom JavaScript code with jsPsych (De Leeuw, 2015). Instructions prior to the experimental trials provided examples of how to interpret the gaze crosshair. Participants were given feedback on a pair of parent–child interaction trials to ensure that they understood the task.

Each experimental trial in the HSP proceeded as follows: participants were shown a video snippet, with a beep occurring at the moment that the original target word was named. After the video, participants were then shown a numbered grid with all 24 toys used in Yu et al. (2021) and asked to make a multiple-choice selection for the object that they

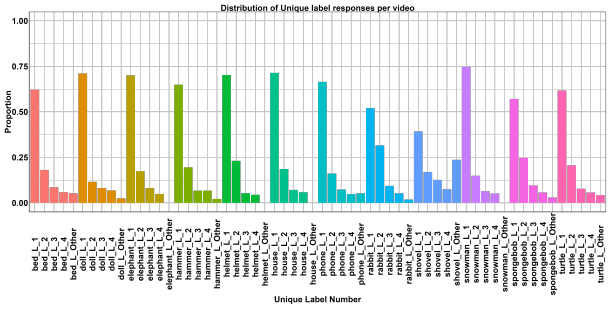
[1] We had further planned to exclude participants based on performance criteria of trial reaction time (faster than 600ms) or overall referent accuracy (less than 10%) as these may indicate lack of attention or noncompliance with the task. However, no participants failed these exclusion measures.



(a)



(b)



(c)

Figure 3: Frequency distribution of unique participant responses per video in the HSP. This is shown at the population-level for both referents and labels in (a), bars indicate standard error. The same distribution split up by target-word is plotted separately for referents (b) and labels (c)

thought the parent was referring to at the moment of the beep. Following their referent guess, participants were asked to guess the exact word that they thought the parent said during the beep (via free-response text box). Instructions encouraged participants to provide their best guess for both referents and labels even if they were not certain.

Results

The two types of participant responses gathered in the egocentric HSP (referent and label guesses) were analyzed to address three core questions: how ambiguous is the input for participants, does this ambiguity reside primarily in the referential or semantic domain, and what social-behavioral factors drive participants' responses.

Distribution of referent and label responses

In order to get an initial assessment of ambiguity, we calculated the average number of unique participant responses per video. The mean number of distinct referent guesses per video is about three and a half, while the mean number of distinct label responses is five. When we ranked each unique response (both referents and labels) per video by frequency and examined the *distribution* of responses, we found that it is highly skewed. As evident in Figure 3, the vast majority of responses to each video are either the first or second most common across participants. Furthermore, this pattern holds for each of the twelve target words in the experiment (see Figures 3b and 3c). It is very rarely the case that a set of videos results in more than two common shared guesses: the top two guesses comprise at least 80% of responses for 90% of all videos. This suggests that while the input can in some ways be considered ambiguous (as we discuss in our analysis of accuracy below), there are strong visual cues guiding learners to consider only an extremely limited number of candidates at any particular naming moment — this is perhaps more closely aligned with the structure of competitors in word learning experiments (e.g. Trueswell, Medina, Hafri, & Gleitman, 2013; Yu & Smith, 2007) than one might assume.

Response Accuracy

Next we analyzed participants' accuracy in recovering the original referents and labels produced by the parents. Figure 4 shows a histogram of response accuracy from individual vignettes. Similar to the attention results reported in (Yu et al., 2021), we also see a bimodal distribution from HSP accuracy: videos tend to either be highly informative (leading participants to the correct referent most of the time) or quite misleading (leading participants consistently to a wrong guess). The non-unimodality of the distribution of referent accuracy per video can be confirmed via Hartigan's dip test (Hartigan & Hartigan, 1985): D's $< .06$, uncorrected p's $< .001$, with modes at 2% and 88% accuracy.

Relationship between label and referent responses

In order to assess the relationship between referential and semantic ambiguity via the HSP, we explored the correlation between the types of responses. We found that the response types on an individual trial are tightly linked: if participants guess the correct referent, then they also guess the exact label with 78% accuracy (compared to only 34% label accuracy on all trials). Naturally, for trials on which participants guess the incorrect referent, they are extremely unlikely to produce the correct label (3%). This pattern is shown in Figure 5a. Analyzing the data at the video- rather than trial-level we also find a strong linear relationship (R^2=.65) between average referent accuracy and average label accuracy (Figure 5b).

For the 22% of trials on which participants guessed the correct referent but did not guess the original label, we categorized incorrect labels in terms of their semantic relation to the intended word. The majority (53%) of these label "errors" may actually be considered synonyms of the word that the parent said. This includes cases like *"bunny"* instead of *"rabbit"* and *"tortoise"* instead of *"turtle."* The next most common categories were using a superordinate term when the parent said something more specific (24%), e.g. *"block"* instead of *"Spongebob"* or *"tool"* instead of *"saw"* or describing the wrong dimension or conceptualization of the referent (20%), e.g. *"girl"* instead of *"doll"* or *"dinosaur"* instead of *"elephant."* The remaining 3% of cases reflect either subordinate terms or a different object all together. Overall, the fact that semantic ambiguity is so strongly reduced in light of identifying the correct referent suggests that referential ambiguity is the primary bottleneck for early word learning.

Which cues drive participant responses?

Given the bimodal distribution over accuracy at the video-level, we explored what factors may cause a vignette to be highly-informative or misleading. Using the median split from Figure 4, we calculated the mean value of various behavioral cues that were present in those videos and therefore might serve to diminish referential ambiguity. Figure 6 shows the proportion of time within each vignette type (high- vs. low-informativity) that the infant was looking at the target referent (Infant-gaze), the proportion of time the infant was holding the target referent (Infant-holding), and the proportion of time the parent was holding the target referent (Parent-holding). We find notably higher rates of all three behavioral cues in the high-informativity videos compared with the low-informativity ones. This trend was confirmed statistically by fitting a logistic regression to predict high vs. low-informativity status of each video, using Infant-gaze ($\beta = 2.28$, $p < .01$), Infant-holding ($\beta = 1.20$, $p < .02$), and Parent-Holding ($\beta = 1.99$, $p < .01$) as independent variables, and a random intercept for each target word.

Regardless of whether participants are able to identify the correct target, we are interested in how these behavioral cues are used by participants to converge to a single referent as shown by the highly skewed distribution in Figure 3. To analyze this, we first calculated the most common referent response (MCR) — which may or may not be the target — from each vignette. We then calculated the most-looked at, and most-held referents based on infants' and parents' behavior.

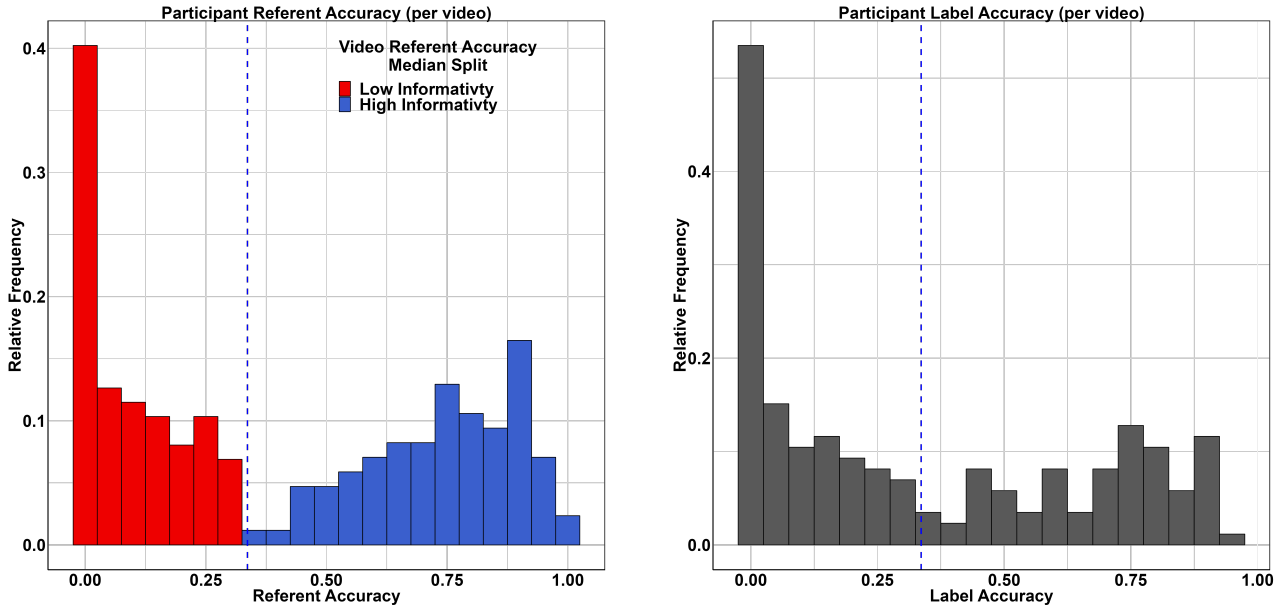


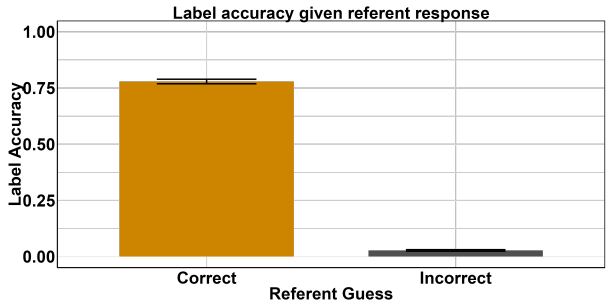
Figure 4: Histogram of participant accuracy with respect to referents (a) and labels (b) for individual vignettes. Referent responses are colored according to median split as this categorization will be used for subsequent analysis in the paper.

Finally, we calculated how well the behavioral cues match with the MCR. For individual cues, the object most-held by the infant was consistent with the participant MCR for 28% of videos, whereas the object most-held by the parent was consistent with MCR for 48% of videos and the MCR had the highest Infant-gaze in 49% of videos (left half of Figure 7). This pattern suggests that participants are using these cues to guess target referents, and in particular that infant-gaze and parent-holding are more important cues than infant-holding. In order to understand how these different cues may interact, we identified the instances wherein two out of the three cues pointed towards the same object (e.g. cases in which infant-gaze and infant-holding both suggested the same referent) and calculated how frequently those cue-pairs matched the MCR (right-half of Figure 7). We find that when Infant-gaze and Parent-holding point towards the same object, then that cued referent matches the MCR 72% of the time. Similarly, the agreement between Infant-holding and Parent-holding is consistent with MCR 71% of the time. Conversely, when Infant-gaze and Infant-holding point to the same object, it only matches with the MCR 44% of the time — slightly lower than Infant-gaze alone. Taken together, the results suggest that both Infant-Gaze and Parent-holding are strong behavioral cues in guiding participant responses while Infant-holding is less informative because it probably provides redundant and less precise information in the presence of Infant-gaze.

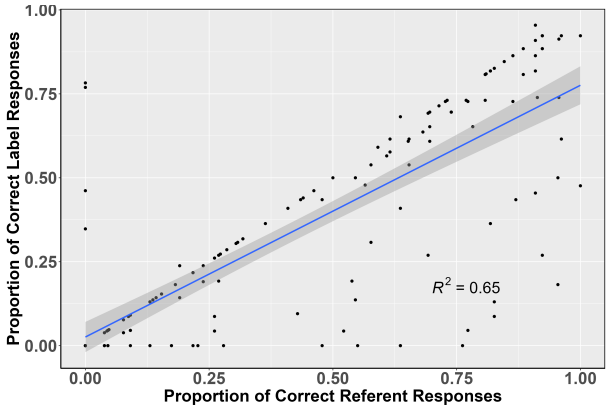
General Discussion

In this paper, we used an extended, egocentric Human Simulation Paradigm (HSP) to quantify the properties of the input available for early word learning. Unlike large-scale studies on the distributional properties of early language input (e.g. Bergelson et al., 2019; Soderstrom, 2007; Yurovsky, Doyle, & Frank, 2016) this paper focused on understanding aspects of input that cannot be accessed on the surface, but only as a result of information processing. This was accomplished by extracting naming instances from a parent-infant toy-play videos corpus (Yu et al., 2021) which included recorded gaze data, playing those videos to adult learners, and gathering their guesses as to the intended referent as well as the original label. We found that snippets from this representative sample were in some ways less ambiguous than one might assume; about 90% of guesses were divided between only two common responses for each video. This is surprisingly consistent with the co-occurring statistics used in common cross-situational learning paradigms which tend to have a small handful of potential referents on the screen at any given time (Yu & Smith, 2007, among others). This trend held true across videos containing all twelve target words in our study.

Consistent with previous findings on the bimodal distribution of infant gaze during toy-play (Yu et al., 2021), we found that participant accuracy across videos also followed a clear bimodal distribution. Vignettes tended to be either highly informative (leading participants to the correct object a majority of the time) or quite misleading (almost never generating a correct response). What's more, our findings suggest that the bottleneck for early word learning resides in referential rather than semantic ambiguity. In the HSP, on trials when the participants guess the correct referent, then they are extremely likely to precisely recover the intended label (and



(a)



(b)

Figure 5: Participant label accuracy given their referent response at the trial-level (a) and the correlation between participants' referent and label response accuracy at the video-level.

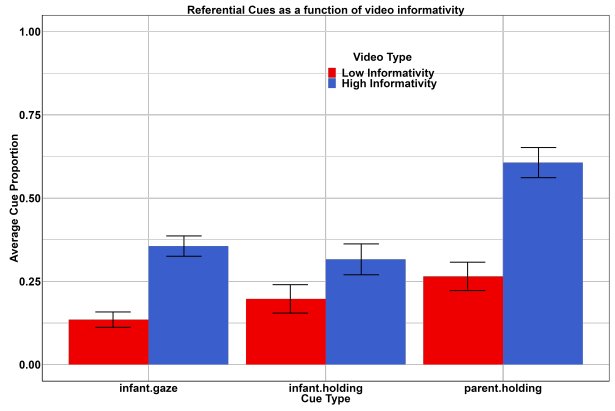


Figure 6: Average proportion of various behavioral cues as a function of video informativity.

thus concept) as well. To put it another way, once participants discover the referent, then the rest of the task (getting the exact meaning) is evidently easier than one might anticipate. Conversely, if one cannot uncover the intended referent then of course there is no hope to recover the right meaning. This may simply be due to the fact that the first words that infants learn are simple, concrete nouns which do not

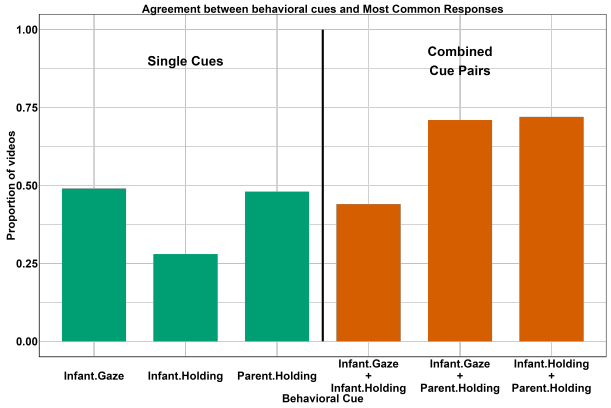


Figure 7: Proportion of videos for which participants' most common referent response (MCR) aligns with the dominant object suggested by various behavioral cues and cue-combinations.

possess the same potential for ambiguity as later acquired, more abstract concepts. This is presumably also aided by factors like the basic-level (Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976) and whole-object (Markman, 1990) biases — participants identify the labels that parents use most frequently since that is the most commonly used label to refer to a target referent in the situation.

Finally, participant responses in the task seems to rely on the same sorts of social-behavioral cues utilized by young learners in the real world; e.g. visual attention and physical holding of objects (Cheung, Hartley, & Monaghan, 2021; Schroer & Yu, 2022). Such behavioral cues, naturally occurring in parent-child social interaction, both predict accuracy at guessing named targets, as well as shared pitfalls for misleading vignettes. It is striking, however, that Infant-gaze (which was directly recorded in the play sessions and displayed on the videos via a brightly-colored crosshair) seemed to play no stronger a role than Parent-holding information (based on its predictive power in cue-pairs). How exactly these cues really compete with one another or are incorporated during learning requires future, more targeted study. For instance, one may experimentally manipulate the presence or absence of cues across videos or conditions (or even artificially shift the crosshair to alternative locations) to see how that affects guesses or learning. Nonetheless, the present study contributes to our understanding of early word learning by beginning to quantify the natural input available for infant learners. A richer understanding of the input, in tandem with theories about the cognitive processes which the input feeds into, will be required for a full understanding of word learning in the real world.

Acknowledgments

We thank members of the Developmental Intelligence Lab for helpful discussion.

References

Bergelson, E., Casillas, M., Soderstrom, M., Seidl, A., Warlaumont, A. S., & Amatuni, A. (2019). What do north american babies hear? a large-scale cross-corpus analysis. *Developmental science*, *22*(1), e12724.

Bergelson, E., & Swingley, D. (2012). At 6–9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences*, *109*(9), 3253–3258.

Cheung, R. W., Hartley, C., & Monaghan, P. (2021). Caregivers use gesture contingently to support word learning. *Developmental Science*, *24*(4), e13098.

Clark, E. V. (1995). *The lexicon in acquisition* (No. 65). Cambridge University Press.

De Leeuw, J. R. (2015). jspsych: A javascript library for creating behavioral experiments in a web browser. *Behavior research methods*, *47*, 1–12.

Dupoux, E. (2018). Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner. *Cognition*, *173*, 43–59.

Fodor, J. A. (1983). *The modularity of mind: An essay on faculty psychology*. MIT press.

Gentner, D. (1982). Why nouns are learned before verbs: Linguistic relativity versus natural partitioning. *BBN report; no. 4854*.

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

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

Gleitman, L., Cassidy, K., Nappa, R., Papafragou, A., & Trueswell, J. C. (2005). Hard words. *Language learning and development*, *1*(1), 23–64.

Hartigan, J. A., & Hartigan, P. M. (1985). The dip test of unimodality. *The annals of Statistics*, 70–84.

Kretch, K. S., Franchak, J. M., & Adolph, K. E. (2014). Crawling and walking infants see the world differently. *Child development*, *85*(4), 1503–1518.

Markman, E. M. (1990). Constraints children place on word meanings. *Cognitive science*, *14*(1), 57–77.

Medina, T. N., Snedeker, J., Trueswell, J. C., & Gleitman, L. (2011). How words can and cannot be learned by observation. *Proceedings of the National Academy of Sciences*, *108*(22), 9014–9019.

Quine, W. V. O. (1960). *Word and object*. MIT press.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive psychology*, *8*(3), 382–439.

Rosenholtz, R. (2016). Capabilities and limitations of peripheral vision. *Annual review of vision science*, *2*, 437–457.

Schroer, S. E., & Yu, C. (2022). Looking is not enough: Multimodal attention supports the real-time learning of new words. *Developmental Science*, e13290.

Smith, L. B., Jayaraman, S., Clerkin, E., & Yu, C. (2018). The developing infant creates a curriculum for statistical learning. *Trends in cognitive sciences*, *22*(4), 325–336.

Smith, L. B., Yu, C., & Pereira, A. F. (2011). Not your mother's view: The dynamics of toddler visual experience. *Developmental science*, *14*(1), 9–17.

Soderstrom, M. (2007). Beyond babytalk: Re-evaluating the nature and content of speech input to preverbal infants. *Developmental Review*, *27*(4), 501–532.

Spelke, E. S., & Kinzler, K. D. (2007). Core knowledge. *Developmental science*, *10*(1), 89–96.

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

Waxman, S. R., & Markow, D. B. (1995). Words as invitations to form categories: Evidence from 12-to 13-month-old infants. *Cognitive psychology*, *29*(3), 257–302.

Xu, F., & Tenenbaum, J. B. (2007). Word learning as bayesian inference. *Psychological review*, *114*(2), 245.

Yang, C. D. (2004). Universal grammar, statistics or both? *Trends in cognitive sciences*, *8*(10), 451–456.

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

Yu, C., Zhang, Y., Slone, L. K., & Smith, L. B. (2021). The infant's view redefines the problem of referential uncertainty in early word learning. *Proceedings of the National Academy of Sciences*, *118*(52), e2107019118.

Yurovsky, D., Doyle, G., & Frank, M. C. (2016). Linguistic input is tuned to children's developmental level. In *Cogsci.*

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.