Partial Word Learning from Referentially Ambiguous Naming Events: Evidence from a Human Simulation Paradigm

Nina Schoener[1] (nina.schoener@uconn.edu)
Peter C. Schoener (peter.c.schoener@gmail.com)
Sara Johnson[1] (sara.2.johnson@uconn.edu)
Sumarga H. Suanda[1] (s.suanda@uconn.edu)
[1] Department of Psychological Sciences, University of Connecticut
406 Babbidge Road, Unit 1020, Storrs, CT, 06269, USA

Abstract

Children learn word meanings from their patterns of usage in their everyday input. This trivial statement is made more interesting by the fact that most patterns of word usage, even around language-learning children, are not particularly good at revealing word meaning. So, do children simply ignore much of their input and learn words primarily from the few instances of usage where word meaning is transparent? Or are there pieces of information in the sea of opaque word usage that would allow children to learn words slowly over time? In an adaptation of Gleitman and colleagues' classic Human Simulation Paradigm (Gillette et al., 1999), the current study explores the kinds of input that contribute to learning. Our data suggest that the answer may depend on how "learning" is assessed.

Keywords: word learning; referential uncertainty; statistical learning; partial knowledge; semantic development

Introduction

Young word learners encounter words in contexts that vary widely in their referential quality. Sometimes words occur in contexts where the intended referent is highly transparent. Most of the time, however, words occur in contexts where the intended referent is ambiguous and difficult to discern from the extralinguistic setting (see Medina et al., 2011; Trueswell et al., 2016). Although there is little disagreement on whether referentially transparent moments play an important role in shaping learning, there is quite a bit of disagreement on whether the referentially ambiguous moments also contribute to the learning process (Gleitman & Trueswell, 2020). Some contend that word learning is driven primarily, or perhaps solely, by the few referentially transparent moments in the input (Gleitman & Trueswell, 2020; Medina et al., 2011; Trueswell et al., 2013) whereas others propose that word learning is shaped by a broader swath of the input, including referentially ambiguous moments (Smith & Yu, 2008; Yu & Smith, 2012). The goal of the current study is to address these questions and ask whether the answer may in part depend on the task one uses to probe word learning.

Much of the research that has examined the role of referential quality on word learning has utilized the human simulation paradigm (HSP), originally developed by Lila Gleitman and colleagues (Gillette et al., 1999). In the HSP, the referential quality of naming events is assessed by putting naïve observers in the shoes of young learners as they experience their parents' naming events. Specifically, these observers watch video vignettes of parent-child interactions in which parents uttered a target word of interest (e.g., "ball"). The entire vignette is muted except for a beep played at the precise moment parents uttered the target word. The observers' task is to guess the identity of the word, and the probability with which observers guess the target word is used as an index of the referential quality of that naming moment (Medina et al., 2011; Cartmill et al., 2013).

Although the HSP approach has been used to address a range of research questions, from the nature of the referential quality of naming events for different word types (Piccin & Waxman, 2007) to the role of naming event quality on individual differences in vocabulary growth (Cartmill et al., 2013), of particular interest for the current study are experimental studies that have specifically examined how events of differing referential quality contribute to learning (Medina et al., 2011). In one study, Medina and colleagues presented adult observers with a set of mostly low referential quality vignettes of the same target word (e.g., "ball") and assessed whether observers could learn the word across vignettes. Interestingly, one of their findings is that when the vignettes were all low in referential quality observers exhibited very little learning and very little improvement across vignettes (see also Yurovsky et al., 2013). Coupled with the finding that observers only learned in conditions where the set contained a vignette high in referential quality, Medina and colleagues' findings raise doubts about whether referentially ambiguous events advance word learning (see also Gleitman & Trueswell, 2020; Trueswell et al., 2016).

The goal of the current study is to revisit the place of referentially ambiguous naming events in learning. What motivates the current study is the possibility that the primary measure of learning used in the HSP (i.e., having to guess the precise identity of the word) is a very high threshold for capturing learning. Indeed, if one considers the common methods of assessing word learning in developmental populations (preferential looking methods, see Hollich et al., 2000; referent selection methods, see Bloom & Markson, 1998; parent report methods, see Fenson, 2007), few require the level of precision in meaning demanded by the HSP. Thus, the current work asks whether the task demands of the HSP may have masked knowledge of word meaning that can be gleaned from repeated exposures to referentially ambiguous events.

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

Several lines of research support this possibility that the role of referentially ambiguous events may have been underestimated by the classic HSP design. First, there are studies that show some learning via referentially ambiguous events even in cross-situational versions of the classic HSP design (see Yurovsky et al., 2013; Zhang et al., 2015). Second, there is the large body of literature illustrating how task and measurement differences can lead to different estimates of word learning (e.g., Bergelson & Aslin, 2017; Carey & Bartlett, 1978; Hendrickson et al., 2015; Horst & Samuelson, 2008). Third, although more commonly discussed in the context of well-defined semantic domains such as color (e.g., Wagner et al., 2013), emotions (e.g., Hoemann et al., 2019), number (e.g., Wynn, 1992), and time (Tillman & Barner, 2015), even the acquisition of more basic concepts like "cup" and "bowl" appear to undergo a protracted learning trajectory (Ameel et al., 2008), highlighting the possibility that there are stages of partial word knowledge that can be shaped via referentially imperfect input. And finally, several word learning scholars have construed word learning as shaped by multiple interconnected but separate systems, including a system for supporting rapid referent identification in the moment (arguably what is captured by the dependent variable in the HSP) and a system of aggregating word knowledge operating on a much larger time scale (see Clerkin & Smith, 2022; Samuelson & McMurray, 2012; Vlach, 2019; Wojcik et al., 2022). Thus, it is possible that although referentially ambiguous events do little for rapid referent identification, they may nonetheless contribute via a separate process.

The current study explores the potential for partial word learning from referentially ambiguous naming events using a modified cross-situational version of the Human Simulation Paradigm. As in previous HSP studies, participants were shown a series of vignettes belonging to a particular target word (e.g., "apple") that had previously been normed to be referentially ambiguous. The vignettes in the current study come from scenes from children's picture books that had contained the target word in its text. Critically, in addition to asking participants to guess the identity of the target word in a free response (FR) test (as is the case in most HSP studies), we asked participants to complete a two-alternative forced choice (2AFC) test where they guessed which of two new vignettes was more likely to have contained the target word. Of particular interest is whether participants succeeded in the 2AFC test even when they failed to correctly identify the target word in the FR test. This would suggest that participants gleaned "partial knowledge" about a word's meaning from referentially ambiguous naming events.

Methods

Participants

Participants were 60 adults (28 Female, 27 Male, five Nonbinary/Other; M_{age} = 28 yrs, SD_{age} = 3.86 yrs) recruited from the US and UK through Prolific (www.prolific.co). Participants received $5 for completing the study. Although all instructions were in written English, participants were not required to be native English speakers.

Target Vignettes

The stimuli for this study consisted of 96 target vignettes and 96 distractor vignettes. Target vignettes were scenes from children's picture books that had contained one of eight target words in its original text. The eight target words were nouns known to be acquired early by English-learning children (Fenson et al., 2007): *apple, bird, book, dog, door, flower, hat, shoe*. For each target word, twelve scenes that had contained that target word were selected from a corpus of over 300 picture books. Each of the twelve scenes came from a different children's picture book.

Distractor Vignettes

For each of the target vignettes, a distractor vignette was selected for the purposes of the two-alternative forced choice test. Distractor vignettes were randomly selected scenes from children's picture books that did not contain the target word in their original text. Distractor vignettes were selected from books that matched the target age range of that from which the target vignettes were selected.

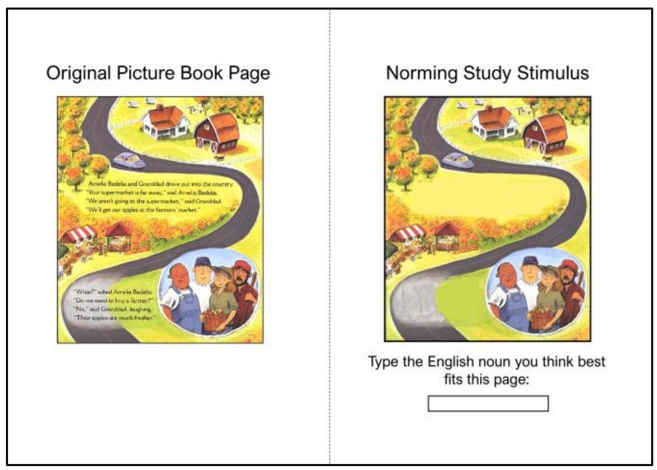

Figure 1: Original picture book page used for target word *apple* (left) and example of norming study trial using the final stimulus created from this page (right).

Vignette Construction and Norming

To create the vignettes, picture book scenes were edited by removing all text on the page and positioned within a 7.5in x 10in PowerPoint slide. All target vignettes had been determined to be "low informative" through prior norming studies. In these norming studies (all hosted online via the Gorilla Experiment Builder, www.gorilla.sc), participants were shown 40 vignettes (each from a unique picture book and each belonging to different target words), one at a time (Anwyl-Irvine et al. 2019). For each vignette, participants entered a noun they thought "fits best on this page" (see Figure 1). Each of the current target vignettes were rated by 16 different participants. Vignette informativity was measured as the proportion of participants who guessed the correct noun. Following Medina and colleagues, .33 was used as the upper threshold for classifying a vignette as low informative. The mean informativity of the current target vignettes was .04 (*SD* = .08, *Range* = 0 - .31).

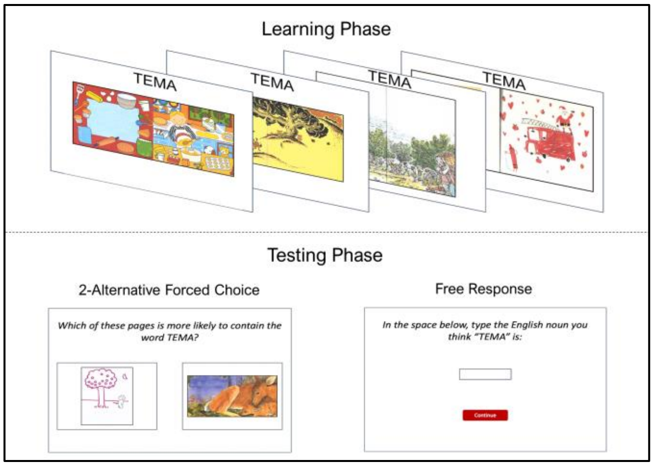

Figure 2: Schematic of learning phase trials (top) and testing phase trials (bottom). The learning phase included nine trials for each of the eight sets and the testing phase included three 2AFC + FR trial pairs per set.

Experimental Design

The experimental design consisted of eight blocks, one block per target word. Each block had a learning phase followed by a testing phase. During the learning phase, participants were shown nine low-informative target vignettes that all originally contained a shared target word in their text. Each target vignette was displayed with a novel word representing the target word that each vignette had in common (Figure 2). Novel words were disyllabic and followed English phonotactic rules, (e.g. "tema"). Participants were not given any explicit task during the learning phase. In the testing phase that immediately followed the learning phase, participants were presented three test trials. Each test trial consisted of a two-alternative force (2AFC) test, followed by a free response (FR) test. In the 2AFC test, participants were shown two new vignettes: a new target vignette that had also originally contained the target word in its text, and a distractor vignette that had not originally contained the target word in its text. Participants were asked to guess which vignette "was more likely to contain the target word". In the FR test, participants were asked to guess the English noun represented by the novel word. For each block, participants completed three 2AFC-FR tests, each with a new target and distractor vignette that had not been seen during the learning phase. Because participants completed eight blocks (for each target word), each participant contributed a total of 24 2AFC and 24 FR tests.

Procedure

Participants gave written consent and completed a brief demographics questionnaire before beginning the experiment. The experiment began with a series of two shortened practice sets consisting of four learning trials and three test trials. The practice sets used target vignettes of English words not used in the experiment proper ("cow", "car", and "cake"). To ensure participants understood the task, after each practice set, participants were shown the correct answers for each 2AFC and FR practice test. The first experiment block started immediately after the two practice sets. Participants typically took 15-20 minutes to complete the entire study and were asked not to take breaks or take any written notes during the experiment.

Results

Free Response Performance

Guesses in the FR test were coded as correct if they shared the same root form of the target word. Thus, guesses that were identical to the target word (e.g., *dog*), a plural form of the target word (e.g., *dogs*), or an infant form of the target word (e.g., *doggy*), we all considered correct. All other responses were considered incorrect. Responses were excluded if participants gave multiple responses including the correct target[1]. Free responses were coded as incorrect in all other cases. Overall, participants guessed the correct target word on .565 of the trials (*SD* = .24).

Alternative Forced Choice Performance

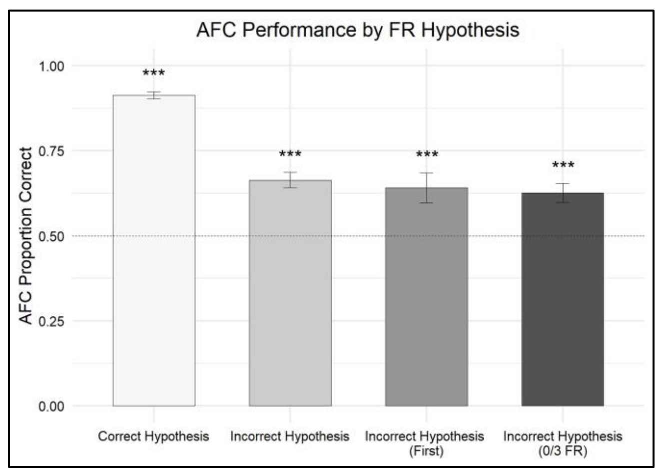

Figure 3: Performance on 2AFC test given (1) a correct hypothesis on the FR test, (2) an incorrect hypothesis on the FR test, (3) an incorrect hypothesis on the FR test, limited to only the first of three test trials for each word, and (4) an incorrect hypothesis on all three test trials. *Note*: *** $p < .001$.

The proportion of 2AFC trials on which participants selected the correct scene was computed. Overall mean proportion correct across participants was .797 (*SD* = .09), a

[1] This was extremely rare in our data, accounting for three out of the 1440 total free response entries (.2%).

proportion that is well above the rate of guessing (.50), $t(59) = 24.86$, $p < .01$, $d = 3.21$. Performance in the 2AFC test was further partitioned as a function of whether the participants guessed the correct target word in the FR prompt that immediately followed the 2AFC trial. Thus, both 2AFC performance with a correct hypothesis and 2AFC performance with an incorrect hypothesis were computed for each participant. Unsurprisingly, participants performed better in the 2AFC test when they had identified the correct target word ($M = .913$, $SD = .081$) than when they had not ($M = .664$, $SD = .176$), $t(58) = 10.87$, $p < .001$, $d = 3.06$, as seen in Figure 3. Importantly, participants' 2AFC performance when they had an incorrect FR guess was still reliably higher than chance ($t(58) = 7.15$, $p < .001$, $d = 0.93$).

Moreover, the finding of significantly greater-than-chance AFC performance despite failing the FR test holds true for six out of the eight individual words tested (Figure 4). These data suggest then even when participants failed to acquire the *precise* meaning of a word, they nonetheless acquired *partial* knowledge that allowed them to succeed in identifying contexts where the word would occur.

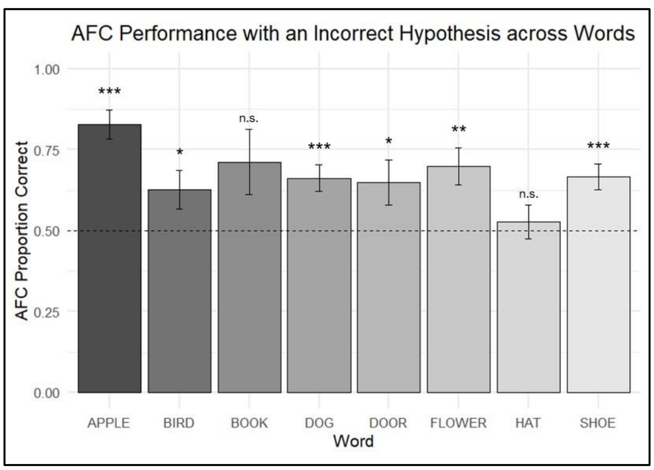

Figure 4: Performance on 2AFC test at the subject level by target word, including only trials for which participants gave an incorrect hypothesis on the free response test. Error bars represent standard errors of the means. *Note*: * $p < .05$, ** $p < .01$, *** $p < .001$.

Two follow-up analyses were conducted to assess the robustness of this partial knowledge effect. First, because there were three separate AFC-FR test pairs for each word (presented consecutively), we tested whether overall performance was the product of learning over the course of the three test trials. Thus, the same key analysis (i.e., performance in AFC trials with an incorrect FR test) was computed but restricted to only the first AFC trial. Partial knowledge on the first AFC trial revealed nearly identical patterns ($M_{first} = .641$, $SD = .33$; $t(57) = 3.19$, $p < .01$, $d = 0.419$) to the analysis of all AFC trials (see Figure 3). Second, because it is possible that participants' FR guesses changed over the course of the three test trials, we tested whether our findings were the product of participants who at *some point* correctly identified the target word. Thus, we reanalyzed the partial knowledge effect but restricted the data to only those words for which participants *never guessed* the correct target word. This follow-up analysis also revealed highly similar patterns to the overall partial knowledge trends reported above ($M_{AFC} = .626$, $SD = .21$; $t(57) = 4.50$, $p < .01$, $d = 0.59$).

Alternative Forced Choice Performance in Relation to Free Response Errors

To explore the nature of this partial learning effect, we examined the relationship between the types of errors participants made in the FR test and their AFC performance for that word. We found that participants' incorrect FR hypotheses were often semantically related to the target word (e.g., the error "*fruit*" for the target word *apple*). Of particular interest was whether the observed partial knowledge effects were driven entirely by such cases where participants offered close-but-incorrect guesses or whether partial knowledge was captured in the AFC test even in cases where participants had less semantically related FR hypotheses.

We approached this by first measuring the semantic similarity between target words and their errors using three different methods. First, we asked a new set of participants ($n = 80$) to rate the errors given by participants in the original study in terms of their similarity to the target word corresponding to that error. For example, if a participant in the original study believed the target word was *sock* when the actual target word was *shoe*, new participants were asked to rate the word *sock* in terms of its semantic similarity to *shoe* on a scale of 1 to 7. Each error from the original study was rated by ten independent raters in terms of its similarity to each of the eight target words. Overall, target-error pairs were rated significantly higher in semantic similarity ($M = 2.92$, $SD = 1.84$) than non-target-error pairs ($M = 1.95$, $SD = .95$; $t(144) = 5.87$, $p < .001$, $d = .527$).

Next, we used two computational models trained on child-directed speech data from CHILDES (as of Dec. 2022; MacWhinney, 2000) to embed our target-error pairs in a vector space representing semantic similarity between words. One model used word2vec (W2V) to determine semantic similarity based on a word's context, (Mikolov et al., 2013). A second model used fastText (FT) to determine semantic similarity based on a similar structure but at the level of morphological subunits (Bojanowski et al., 2016). Across both models[2], as with the human ratings, we found that target-error pairs were higher in semantic similarity ($M_{w2v} = 2.81$, $SD = 1.05$; $M_{ft} = 3.88$, $SD = 0.93$) than non-target-error pairs ($M_{w2v} = 2.25$, $SD = 0.92$; $M_{ft} = 3.32$, $SD = 1.01$) in both the W2V model and the FT model ($t(174) = 4.66$, $p_{w2v} < .001$; $t(195) = 6.28$, $p_{ft} < .001$).

[2] For ease of comparison to the Human Ratings data, similarity scores for the models were converted to a 1-to-7 scale.

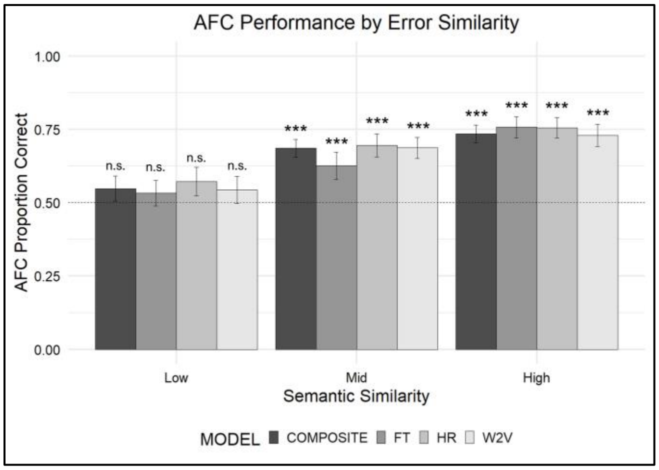

Figure 5: Performance on 2AFC test as a function of the semantic similarity of their FR errors. Data is shown as a function of semantic similarity tertiles across all estimates of similarity: composite, human ratings (HR), fastText embedding model (FT), and word2vec embedding model (W2V). *Note*: *** $p < .001$.

Figure 5 depicts 2AFC test performance as a function of the semantic similarity of FR errors, as well as a composite score that averaged the ratings across the three methods[3]. As the figure depicts, the semantic similarity of errors mattered for AFC performance. Across all estimates of semantic similarity, errors that were highly similar to the target word (based on tertiary split) led to higher AFC performance than errors that were low in similarity to the target word (highest *p*-value across similarity estimates < .01).

Interestingly, however, participants appeared to perform above chance even when their errors were not highly semantic similar to the target. Across all measures, errors that were moderately similar led to AFC performance that was above chance (highest *p*-value across similarity estimates < .001). For all models, errors that had low semantic similarity to the target were associated with performance around chance levels at the AFC task (smallest *p*-value across similarity estimates = .15). Overall, the robust effect of above-chance AFC performance given moderately similar errors[4] does highlight that highly similar hypotheses were sufficient, but not necessary for success in the AFC test.

Discussion

Understanding the kinds of input that shape children's word learning has broad implications, from a better understanding of the mechanisms that shape learning (see Gleitman & Trueswell, 2020; Yu & Smith, 2012) to developing best practices for input-based interventions (see Masek et al., 2021; Rowe & Snow, 2020). Previous research on whether word learners can leverage referentially ambiguous input for learning has yielded mixed results. A number of studies have shown robust learning in cross-situational word learning paradigms where the referentially ambiguous input takes the form of a small set of novel words and novel objects (e.g., Yu & Smith, 2007; see Smith et al., 2014 for review). Others have shown, however, that learners struggle when the referentially ambiguous input takes the form of naming events in the real-world (see Cartmill et al., 2013; Medina et al., 2011; but see Yurovsky et al., 2013). The current study sheds new light on the potential for referentially ambiguous input to move the needle of learning. In an adaptation of Gleitman, Trueswell, and colleagues' Human Simulation Paradigm, the current study reveals that when referentially ambiguous input fails to yield learning of the *exact* word meanings, it often yields partial knowledge that brings learners semantically close and allows them to succeed in some tests of word knowledge.

This notion that word learners develop partial knowledge of word meanings en route to acquiring a new word is neither novel nor unintuitive (see Carey & Bartlett, 1978). Thus, the contribution of the current work is not *whether* partial knowledge can be acquired, but rather the kinds of input from which that knowledge can be acquired. Few studies on partial word learning have focused on the kinds of observational contexts that shape it (see Schwanenflugel et al., 1997, however, for work in the context of partial learning word meaning via reading). Those that have explored the processes of partial word-referent learning have used simplified scenes as their input (see Figure 6; Dautriche & Chemla, 2014; Yurovsky & Frank, 2015; Yurovsky et al., 2014). Although the current work does not take its input from observations of parent-child interactions, as others have (e.g., Medina et al., 2011), the input did consist of complex, multifaceted scenes that represent one type of input that many children experience the world over (see Hudson Kam & Matthewson, 2017).

The exact mechanisms by which participants extracted and deployed partial word knowledge in the current study is unclear. There are at least two possible explanations for the observed success in the Alternative Forced Choice (AFC) trials on the one hand and failure in the Free Response (FR) task on the other.

One is consistent with a "propose-but-verify" model of word-referent learning, in which word learners obtain and maintain a single hypothesis at any point during learning (Trueswell et al., 2013). Under this explanation, participants' single hypotheses may have been insufficient to meet the high threshold requirement of the FR test (e.g., the hypothesis "fruit" is incorrect for the word "apple" because they mean different things), but good enough to meet the lower threshold for success in the AFC test (e.g., the hypothesis "fruit" is likely, on average, a better match for scenes that do contain the word "apple" than for scenes that do not contain

[3] Ratings across similarity estimates were positively correlated (smallest $p < .001$). Unsurprisingly, the strongest correlation was between W2V and FT ($r = .72$); the weakest was between HR and FT ($r = .36$).

[4] Sample mid word-error pairings were: door-leaf, hat-green, and flower-family

the word "apple"). Our findings from the human- and model-based semantic ratings showing that semantically close FR errors were more likely to yield better AFC performance are entirely consistent with this view.

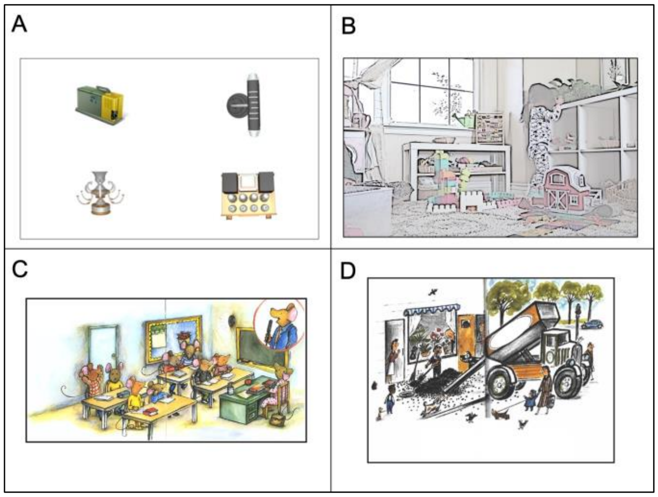

Figure 6. Scene complexity across common paradigms. (A) image representative of cross-situational word learning paradigms; (B) a static image of an at-home video of parent-child interactions[5] (C-D) picture book scenes used in the current study for the words "apple" (C) and "dog" (D)

An alternative explanation is one based on associative learning principles where participants extract and maintain in memory more information from the input than a single hypothesis about word meaning (Yu & Smith, 2012; Knabe & Vlach, 2022; Zettersten et al., 2017). Under this model, the dissociation in performance between the FR and AFC tasks may be due to the fact that although the FR task forced participants to funnel their knowledge into a single hypothesis, the AFC task allowed participants to recruit further knowledge acquired during learning. Although the positive correlation between error semantic similarity and AFC performance described above is by no means inconsistent with this model, another observation in the current study seems consistent with the associative learning approach. That is, identifying a semantically similar error did not appear necessary to succeed in the AFC task. That is, even errors that were only moderately similar in meaning to the target word (e.g., the error "child" for the target word "shoe", or "color" for "apple") led to successful AFC performance. Although by no means definitive[6], these results raise the possibility participants may have recruited knowledge separate from their hypotheses in the AFC test.

In addition to exploring further the nature of the underlying mechanisms, several other issues raised by the current findings are worthy of future investigations. First, although the current study utilized vignettes that are matched in degree of ambiguity as previous studies employing video vignettes (Medina et al., 2011), it is entirely possible that the kinds of ambiguity present in children's picture books is distinct from the ambiguity in child-directed speech, and more amenable to partial knowledge acquisition. Thus, it would be important to examine whether the kinds of partial learning observed in the current study extend to vignettes from child-directed speech. Second, there are features of the current study's vignettes (e.g., none of the vignettes involved absent reference) and paradigm (e.g., vignettes followed a massed presentation schedule) that are likely to have contributed to the observed partial learning. Although the ways in which such features also appear in children's actual word learning input is a continued matter of investigation, these methodological decisions nonetheless raise the question of whether the current findings would be observed with less favorable stimuli or less favorable learning conditions. Finally, like many other HSP studies, the current work used mature adult cognitive systems as models for children's word learning (see Gillette et al., 1999; Medina et al., 2011; Yurovsky et al., 2013; Zhang & Yu, 2021). Although some have revealed similarities in how adults and children perform in the HSP (e.g., Piccin & Waxman, 2007), it would nonetheless be critical to ask whether, like the adults in the current study, young children can extract and deploy partial knowledge from referentially ambiguous input.

Conclusion

Inferences about both the inputs that contribute to learning and the mechanisms that drive it are inherently linked to how learning is assessed. The current data underscore this by demonstrating that our conclusions about whether referentially ambiguous events contribute to word learning depend on the assessment method and its specific threshold for learning. If a participant succeeds at the Alternative Forced Choice test despite failing at the Free Response test, they can still be considered to have learned *something* about the meaning of the target word, even if not its precise meaning. Although some may find that we grant learning status to participants who passed the AFC test but failed the FR test, certain facts and findings about the measurement of lexical development lend some credence to this conclusion. First, many of the methods used to grant learning status to children (e.g., the Peabody Picture Vocabulary Test; Dunn, 2019) are closer in form to our AFC test than they are to our FR test. Additionally, a growing body of research is spotlighting how people differ in their representations of even commonly used words (see Wang & Bi, 2021), and that such differences may be larger in development (see Ameel et al., 2008). Thus, there may be more to be gained from a broader conception of learning, than a more restrictive one.

[5] Source: https://www.youtube.com/watch?v=YeZ8XP0HtTs

[6] It is possible that the AFC target scenes were more consistent not only with semantically-similar errors, but also with semantically dissimilar errors. If this were the case, then even the semantically-distant above-chance performance could be consistent with a single-hypothesis account.

Acknowledgements

We thank the members of the UConn Communication and Development Lab for their assistance. This research was supported by the James S. McDonnell Foundation (JSMF 220020549) and The National Institutes of Health (R00-HD082358).

References

Bergelson, E., & Aslin, R. N. (2017). Nature and origins of the lexicon in 6-mo-olds. *Proceedings of the National Academy of Sciences, 114*(49), 12916–12921. https://doi.org/10.1073/pnas.1712966114

Bloom, P., & Markson, L. (1998). Capacities underlying word learning. *Trends in Cognitive Sciences, 2*(2), 67–73. https://doi.org/10.1016/S1364-6613(98)01121-8

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2016). Enriching word vectors with subword information. ArXiv.

Carey, S., & Bartlett, E. (1978). Acquiring a single new word. *Proceedings of the Stanford Child Language Conference, 15*, 17-29.

Cartmill, E. A., Armstrong, B. F., Gleitman, L. R., Goldin-Meadow, S., Medina, T. N., & Trueswell, J. C. (2013). Quality of early parent input predicts child vocabulary 3 years later. *Proceedings of the National Academy of Sciences, 110*(28), 11278–11283. https://doi.org/10.1073/pnas.1309518110

Clerkin, E. M., & Smith, L. B. (2022). Real-world statistics at two timescales and a mechanism for infant learning of object names. *Proceedings of the National Academy of Sciences, 119*(18), e2123239119. https://doi.org/10.1073/pnas.2123239119

Dautriche, I., & Chemla, E. (2014). Cross-situational word learning in the right situations. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*(3), 892–903. https://doi.org/10.1037/a0035657

Fenson, L., Marchman, V. A., Thal, D. J., Dale, P. S., Reznick, J. S., & Bates, E. (2007). *MacArthur-Bates Communicative Development Inventories: User's Guide and Technical Manual. Second Edition.* Baltimore, MD: Brookes Publishing Co.

Gillette, J., Gleitman, H., Gleitman, L., & Lederer, A. (1999). Human simulations of vocabulary learning. *Cognition, 73*(2), 135–176. https://doi.org/10.1016/S0010-0277(99)00036-0

Gleitman, L. R., & Trueswell, J. C. (2020). Easy words: Reference resolution in a malevolent referent world. *Topics in Cognitive Science, 12*(1), 22–47. https://doi.org/10.1111/tops.12352

Hendrickson, K., Mitsven, S., Poulin-Dubois, D., Zesiger, P., & Friend, M. (2015). Looking and touching: What extant approaches reveal about the structure of early word knowledge. *Developmental Science, 18*(5), 723–735. https://doi.org/10.1111/desc.12250

Hoemann, K., Xu, F., & Barrett, L. F. (2019). Emotion words, emotion concepts, and emotional development in children: A constructionist hypothesis. *Developmental Psychology, 55*(9), 1830–1849. https://doi.org/10.1037/dev0000686

Hollich, G. J., Hirsh-Pasek, K., & Golinkoff, R. M. (Eds.). (2000). Breaking the language barrier: An emergentist coalition model for the origins of word learning. Blackwell.

Horst, J. S., & Samuelson, L. K. (2008). Fast mapping but poor retention by 24-month-old infants. *Infancy, 13*(2), 128–157. https://doi.org/10.1080/15250000701795598

Hudson Kam, C. L., & Matthewson, L. (2017). Introducing the infant bookreading database(Ibdb). *Journal of Child Language, 44*(6), 1289–1308. https://doi.org/10.1017/S0305000916000490

Knabe, M. L., & Vlach, H. A. (2023). Not all is forgotten: Children's associative matrices for features of a word learning episode. *Developmental Science*, *26*(2). https://doi.org/10.1111/desc.13291

Masek, L. R., Ramirez, A. G., McMillan, B. T. M., Hirsh-Pasek, K., & Golinkoff, R. M. (2021). Beyond counting words: A paradigm shift for the study of language acquisition. *Child Development Perspectives, 15*, 274-280.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk* (3rd ed). Lawrence Erlbaum.

McMurray, B., Horst, J. S., & Samuelson, L. K. (2012). Word learning emerges from the interaction of online referent selection and slow associative learning. *Psychological Review, 119*(4), 831–877. https://doi.org/10.1037/a0029872

Medina, T. N., Snedeker, J., Trueswell, J. C., & Gleitman, L. R. (2011). How words can and cannot be learned by observation. *Proceedings of the National Academy of Sciences, 108*(22), 9014–9019. https://doi.org/10.1073/pnas.1105040108

Mikolov, T., Chen, K., Carrado, G. and Dean, J., 2013. Efficient Estimation of Word Representations in Vector Space. 1st ed.

Piccin, T. B., & Waxman, S. R. (2007). Why nouns trump verbs in word learning: New evidence from children and adults in the human simulation paradigm. *Language Learning and Development, 3*(4), 295–323. https://doi.org/10.1080/15475440701377535

Rowe, M. L., & Snow, C. E. (2020). Analyzing input quality along three dimensions: Interactive, linguistic, and conceptual. *Journal of Child Language, 47*(1), 5-21.

Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition, 106*(3), 1558–1568. https://doi.org/10.1016/j.cognition.2007.06.010

Smith, L. B., Suanda, S. H., & Yu, C. (2014). The unrealized promise of infant statistical word–referent learning. *Trends in Cognitive Sciences, 18*(5), 251–258. https://doi.org/10.1016/j.tics.2014.02.007

Schwanenflugel, P. J., Stahl, S. A., & McFalls, E. L. (1997). Partial word knowledge and vocabulary growth during reading comprehension. *Journal of Literacy Research, 29*(4), 531–553. https://doi.org/10.1080/10862969709547973

Tillman, K. A., & Barner, D. (2015). Learning the language of time: Children's acquisition of duration words.

Cognitive Psychology, *78*, 57–77. https://doi.org/10.1016/j.cogpsych.2015.03.001

Trueswell, J. C., Lin, Y., Armstrong, B., Cartmill, E. A., Goldin-Meadow, S., & Gleitman, L. R. (2016). Perceiving referential intent: Dynamics of reference in natural parent–child interactions. *Cognition*, *148*, 117–135. https://doi.org/10.1016/j.cognition.2015.11.002

Trueswell, J. C., Medina, T. N., Hafri, A., & Gleitman, L. R. (2013). Propose but verify: Fast mapping meets cross-situational word learning. *Cognitive Psychology*, *66*(1), 126–156. https://doi.org/10.1016/j.cogpsych.2012.10.001

Vlach, H. A. (2019). Learning to remember words: Memory constraints as double-edged sword mechanisms of language development. *Child Development Perspectives*, *13*(3), 159–165. https://doi.org/10.1111/cdep.12337

Wagner, K., Dobkins, K., & Barner, D. (2013). Slow mapping: Color word learning as a gradual inductive process. *Cognition*, *127*(3), 307–317. https://doi.org/10.1016/j.cognition.2013.01.010

Wojcik, E. H., Zettersten, M., & Benitez, V. L. (2022). The map trap: Why and how word learning research should move beyond mapping. *WIREs Cognitive Science*, *13*(4). https://doi.org/10.1002/wcs.1596

Wynn, K. (1992). Children's acquisition of the number words and the counting system. *Cognitive Psychology*, *24*(2), 220–251. https://doi.org/10.1016/0010-0285(92)90008-P

Yu, C., & Smith, L. B. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*, *18*(5), 414–420. https://doi.org/10.1111/j.1467-9280.2007.01915.x

Yu, C., & Smith, L. B. (2012). Embodied attention and word learning by toddlers. *Cognition*, *125*(2), 244–262. https://doi.org/10.1016/j.cognition.2012.06.016

Yu, C., & Smith, L. B. (2012). Modeling cross-situational word-referent learning: Prior questions. *Psychological Review*, *119*, 21-39.

Yurovsky, D., & Frank, M. C. (2015). An integrative account of constraints on cross-situational learning. *Cognition*, *145*, 53-62. https://doi.org/10.1016/j.cognition.2015.07.013

Yurovsky, D., Fricker, D. C., Yu, C., & Smith, L. B. (2014). The role of partial knowledge in statistical word learning. *Psychonomic Bulletin & Review*, *21*(1), 1–22. https://doi.org/10.3758/s13423-013-0443-y

Yurovsky, D., Smith, L. B., & Yu, C. (2013). Statistical word learning at scale: The baby's view is better. *Developmental Science*. https://doi.org/10.1111/desc.12036

Zettersten, M., Wojcik, E., Benitez, V. L., & Saffran, J. (2018). The company objects keep: Linking referents together during cross-situational word learning. *Journal of Memory and Language*, *99*, 62–73. https://doi.org/10.1016/j.jml.2017.11.001

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). https://doi.org/10.1111/cogs.13010