Referential cues modulate attention and memory during cross-situational word learning

Kyle MacDonald
kyle.macdonald@stanford.edu
Department of Psychology
Stanford University

Daniel Yurovsky
yurovsky@stanford.edu
Department of Psychology
Stanford University

Michael C. Frank
mcfrank@stanford.edu
Department of Psychology
Stanford University

Abstract

Tracking word-object co-occurrence statistics can reduce referential uncertainty during word learning. But human learners are constrained by limits on attention and memory, and therefore must store a subset of the information available—how do they select what information to store? We hypothesize that the presence of referential cues like eye gaze guides how learners allocate their attention. In three large-scale experiments with adults, we test how the presence of referential cues affects cross-situational word learning. Referential cues shift learners away from multiple hypothesis tracking towards storing only a single hypothesis (Experiments 1 and 2). In addition, learners are sensitive to the reliability of a referential cue and when it is less reliable, they are less likely to use the cue and more likely to store multiple hypotheses (Experiment 3). Together, the data suggest a rational tradeoff: In conditions of greater uncertainty, learners tend to store a broader range of information.

Keywords: statistical learning, word learning, referential cues, resource rationality

Introduction

Words are powerful tools that allow speakers to rapidly convey meaning. We focus here on the task of mapping concrete nouns to objects, as opposed to other substantive inferential problems such as word segmentation and generalization. To make such a mapping, learners must solve the core problem of referential uncertainty (Quine, 1960): that a speaker's utterance could refer to many possible objects in the visual scene, to parts of of those objects, or even to something that is not present. How do learners infer word meanings from data with this kind of uncertainty?

Statistical learning theories offer a solution to this learning problem by aggregating cross-situational statistics across labeling events to identify underlying word meanings. Recent experimental work shows that both adults and young infants can use word-object co-occurrence statistics to learn words from individually ambiguous naming events (L. Smith & Yu, 2008; Vouloumanos, 2008). For example, L. Smith and Yu (2008) taught 12-month-olds three novel words simply by repeating consistent novel word-object pairings across 10 ambiguous exposure trials. Moreover, recent computational models suggest that cross-situational learning can scale up to learn adult-sized lexicons, even under conditions of considerable referential uncertainty (K. Smith, Smith, & Blythe, 2011).

Although all learning models agree that the input is the co-occurrence between words and objects, they disagree about how closely learners approximate the input distribution. Some theories hold that we accumulate graded, statistical evidence about multiple referents for each word (McMurray, Horst, & Samuelson, 2012), while others argue that we track only a single candidate referent (Trueswell, Medina, Hafri, & Gleitman, 2013). Recent experimental and modeling work suggests an integrative explanation: that learners store both a strong single hypothesis and a set of weaker alternative hypotheses, with the strength of the alternatives modulated by the number of referents present during learning (Yurovsky & Frank, under review). Under Yurovsky and Frank (under review)'s model, learners allocate a fixed amount of their attention to one hypothesis, and the rest gets distributed evenly among the remaining alternatives. As the set of alternatives grows, the amount allocated to each object approaches zero.

This framework raises the interesting question of whether learners might be sensitive to the quality of the learning context and use this information to adaptively allocate their fixed cognitive resources. This characterization fits well with recent modeling and experimental work that attempts to offer resource-rational explanations of higher cognition (Lieder, Goodman, & Griffiths, 2014). For example, Vul, Goodman, Griffiths, and Tenenbaum (2014) showed that as time-pressure increased in a decision-making task, participants were more likely to show behavior consistent with a less cognitively challenging strategy of matching, rather than with the globally optimal strategy. Are there comparable aspects of the word learning context that might shift learners' allocation of cognitive resources?

Here we consider the hypothesis that referential cues (e.g., eye gaze and pointing) modulate learners' resource allocation by providing evidence about the speaker's intended meaning. Social-pragmatic theories of language learning have emphasized referential cues as critical for early word learning (Clark, 2009). Experimental work shows that even children as young as 16 months are sophisticated intention-readers, preferring to map novel words to objects that are the target of a speaker's gaze and not their own (Baldwin, 1993). And in naturalistic observations, learners tend to retain labels that are accompanied with clear referential cues that are concurrent with visual access (Yu & Smith, 2012). Together, the evidence suggests that referential cues could help learners by allowing for efficient allocation of limited attention to the relevant statistics in the input.

In the current set of studies, we test the effect of referential cues on the number of representations stored during cross-situational word learning. In Experiment 1, we manipulate the presence of a valid referential cue, a speaker's eye gaze, at different levels of attention and memory demands. At all levels of difficulty, learners tracked a strong single hypothesis, but learners were less likely to track multiple word-object links when referential cues were present. In Experiment 2, we replicate the findings from Experiment 1 with a more ecologically valid stimulus set. In Experiment 3, we show that reducing the reliability of the referential cue increases learners multiple hypothesis tracking. Together, the data suggest that learners adaptively allocate attention and store representations with different levels of fidelity depending on the amount of referential uncertainty present during learning.

Experiment 1

We set out to test the effects of referential cues on cross-situational learning at different levels of attention and memory demands. Participants saw a series of ambiguous exposure trials that consisted of a set of novel objects (either 2, 4, 6, or 8) and an image of a schematic, female interlocutor. On each trial they heard one novel word that was either paired with an eye gaze cue or not, and were asked to make guesses about which object went with each word. In subsequent test trials, participants heard the novel word again after different numbers of intervening trials (0, 1, 3, and 7), this time paired with a new set of novel objects. Importantly, test trials were contingent upon participants' selection during exposure such that one of the objects in the set was either the participant's initial guess (Same trials) or one of the objects that was *not* the initial guess (Switch trials). While both single and multiple referent trackers could succeed on Same trials, only participants who encoded multiple word-object links during their first encounter could succeed on Switch trials. This provides a direct test of whether learners track multiple alternatives and if these representations are influenced by the presence of referential cues.

Methods

Participants This experiment was posted to Amazon Mechanical Turk as a set of Human Intelligence Tasks (HITs) to be completed only by participants with US IP addresses and an approval rate above 95%. Each HIT paid 30 cents. Approximately 50-130 HITs were posted for each of the 32 conditions (4 referents X 4 intervals X 2 gaze conditions) for total of approximately 2400 paid HITs. If a participant completed the experiment more than once, he or she was paid each time but only data from the first HITs completion was included in the final data set. In addition, data was excluded from the final sample if participants did not give correct answers for familiar trials (5 HITs excluded).

Stimuli Figure 1 shows stimuli used in Experiment 1. These stimuli consisted of black and white pictures of familiar and novel objects drawn from the set of 140 first used in Kanwisher, Woods, Iacoboni, and Mazziotta (1997), a schematic drawing of a human interlocutor, and audio recordings of familiar and novel words. Familiar words consisted of the labels for the familiar objects as produced by AT&T Natural Voices^{TM}(voice: Crystal). Novel words were 1–3 syllable

Figure 1: Experimental stimuli from Experiment 1 (schematic) and Experiment 2 (live action).

pseudowords obeying the rules of English phonotactics produced using the same speech synthesizer. A schematic drawing of a human speaker was chosen for ease of manipulating the direction of eye gaze, the referential cue of interest in this study (see Figure 1).

Design and Procedure Participants were exposed to a series of 16 trials (8 exposure, and 8 test) in which they heard a speaker say one novel word, saw a set of novel objects, and were asked to guess which object went with the word. After a written explanation of the task, participants completed four practice trials that consisted of familiar words and objects. These trials also served to screen for participants who did not have their audio enabled or who were not attending to the task.

After the practice trials, participants were informed that they would now hear novel words, and see novel objects, and that they should continue selecting the correct referent for each word. Participants heard eight novel words twice, one exposure and one test trial for each word. Four of the test trials were *Same* trials in which the object that participants selected on the exposure trial appeared again amongst a set of new objects. The other four were *Switch* trials in which one of the objects in the set was selected randomly from the objects that the participant did not select on the previous exposure trial. All other objects were completely novel on each trial.

Participants were randomly assigned to see either 2, 4, 6, or 8 referents on each trial and to have either 0, 1, 3, or 7 trials in between exposure and test. Participants were also assigned to either the Gaze or No-gaze condition. In the Gaze condition, eye gaze was directed towards one of the objects on exposure trials; in the No-gaze condition, eye gaze was always directed straight ahead. On test trials, eye gaze was never informative. To indicate that participants' selections had been registered, a red dashed box appeared around the object they selected for 1 second after their click was received. This box appeared around the selected object whether or not it was the "correct" referent.

Results and Discussion

Exposure trials To ensure that our referential cue manipulation was effective we compared participant's performance

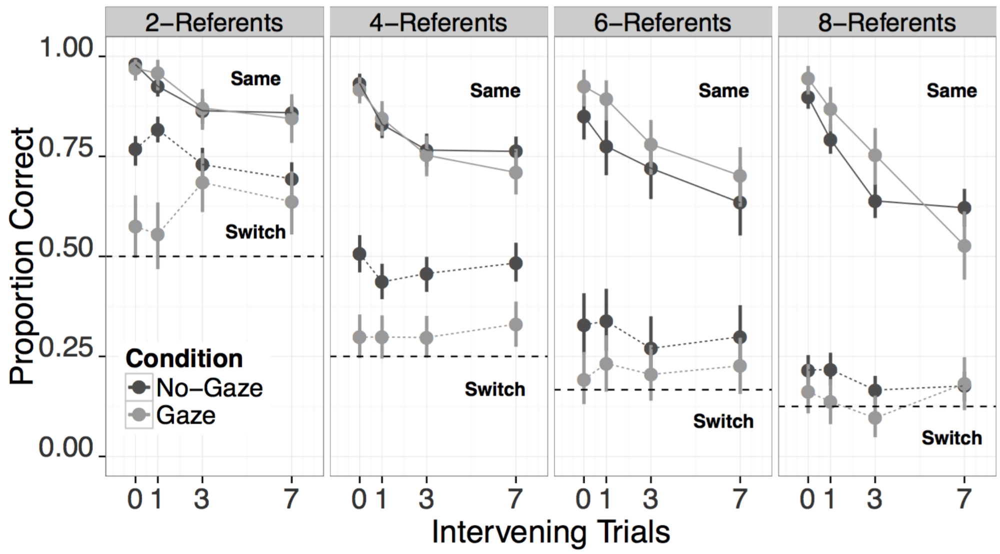

Figure 2: Accuracy on test trials in Experiment 1 for both trial types (Same and Switch) and experimental conditions (Gaze and No-gaze). Each datapoint represents approximately 35-130 participants. Error bars indicate 95% confidence intervals computed by non-parametric bootstrap.

on Exposure trials in the Gaze condition against the distribution expected if participants were selecting randomly (defined by a Binomial distribution with four trials and a probability of success of $\frac{1}{\#Referents}$). In all conditions, participants' responses differed from those expected by chance, exact binomial p(two-tailed) $< .001$, suggesting that eye gaze effectively directed participants' attention to the target referent.

We also analyzed participants' response times on exposure trials, which were self-paced and thus a proxy for attention allocated to the referents on the screen. We fit a linear mixed effects model to response times on exposure trials.[1] We found a significant main effect of referents ($\beta = 806.88$, $p < .001$) with slower responses as the number of referents increased. We also found a significant two-way interaction between condition and number of referents ($\beta = -517.36$, $p < .001$) such that responses were faster in the gaze condition, especially as the number of referents increased.

Test trials To analyze performance on Test trials, we compared the distribution of correct responses made by each participant to the distribution expected if participants were selecting randomly. Figure 2 shows participants' accuracies in identifying the referent of each word in all conditions for both kinds of trials (Same and Switch) and in each referential cue condition (Gaze and No-gaze). We replicate the finding from Yurovsky and Frank (under review): at all Referent and Interval levels, both for Same and for Switch trials, participants' responses differed from those expected by chance (smallest $\chi^2(4) = 12.07, p < .01$). Participants' success on Switch trials provides direct evidence that learners encoded more than a single hypothesis in ambiguous word learning situations, even under high attentional and memory demands.

To quantify the effect of each factor on the likelihood of a correct response, we fit a mixed-effects logistic regression model to the full dataset.[2] We found significant main effects of number of referents ($\beta = -0.66$, $p < .001$) and interval ($\beta = -0.48$, $p < .001$), such that as each of these factors increased, accuracy on test trials decreased. We also found significant main effects of trial type ($\beta = -1.29$, $p < .001$), with worse performance on switch trials.

Next we examined the interactions between each factor. There were significant two-way interactions between trial type and interval ($\beta = 0.31$, $p < .001$) and trial type and number of referents ($\beta = -0.69$, $p < .001$) such that the interval between exposure and test affected same trials more than switch trials, and the number of referents affected switch trials more than same trials. The two-way interaction between trial type and gaze condition was not significant in this model, but trended in the correct direction ($\beta = -0.52$, $p = 0.13$).

The interaction between trial type and gaze condition is the important test of our hypothesis because it shows that the presence of a referential cue reduced learners multiple hypothesis tracking. But we would only expect to see the effect of referential cues on test trials if participants were actually using the gaze cue on exposure trials. So we fit a mixed-effects logistic regression to a filtered dataset, removing those

[1] All mixed-effects models were fit using the lme4 package in R. The model was specified as follows: `RT ~ Gaze-condition × Log(Interval) × Log(Referents) + (Trial Type | subject)`.

[2] The model specification was as follows: `Correct ~ Trial Type × Gaze-condition × Log(Interval) × Log(Referents) + (Trial Type | subject)`.

participants who were not reliably using the gaze cue on exposure trials. This filter removed 90 participants who selected the target of eye gaze at below chance levels on exposure trials. The analysis of the filtered dataset showed a reliable interaction between trial type and gaze condition ($\beta = -0.79$, $p < .05$), suggesting that when learners used the gaze cue, they were less likely to track multiple word-object links.

Taken together, the response time and accuracy analyses provide evidence that learners use of a referential cue modulated their attention during learning, thus making them less likely to track multiple word-object links. Interestingly, we did not see strong evidence that reduced tracking of alternatives resulted in a boost to participants' performance on same trials. This finding suggests that the limitations on same trials may be different than those regulating the distribution of attention on switch trials, since the presence of a referential cue selectively reduced learners tracking of alternatives but did not lead learners to form a stronger memory of their single candidate hypothesis.

Experiment 2

In Experiment 2, we attempt to replicate the findings from Experiment 1 using a more ecologically valid stimuli set. To move closer to a real word learning context, we replaced the static, schematic gaze cue with a live actress and introduced a within-subjects design where each participant saw both gaze and no-gaze exposure trials. We selected a subset of conditions, testing only the 4-referent display with 0 and 3 intervening trials as between-subjects manipulations. Our goals were to replicate the effect of referential cues on learners' multiple hypothesis tracking, and to test whether increasing the ecological validity of the cue would result in a boost to the strength of learners' single candidate hypothesis.

Methods

Participants Participant recruitment and inclusionary/exclusionary criteria were identical to those of Experiment 1 (excluded 36 HITs). 100 HITs were posted for each condition (1 referent X 2 intervals X 2 gaze conditions) for total of 400 paid HITs.

Stimuli Audio and picture stimuli were identical to Experiment 1. The referential cue in the gaze condition was a film of a live actress (see Figure 1). On each exposure trial, the actress looked out at the participant with a neutral expression, smiled, and then turned to look at one of the four images on the screen. She maintained her gaze for 3 seconds before returning to the center. On test trials, she looked straight ahead for the duration of the trial.

Design and Procedure Procedures were identical to those of Experiment 1. The major design change was a within-subjects manipulation of the gaze cue. That is, participants saw exposure trials with and without eye gaze. The experiment consisted of 32 trials broken down into 2 blocks of 16 trials. Each block consisted of 8 exposure trials and 8 test trials (4 same test trials and 4 switch test trials), and contained only gaze or no-gaze exposure trials. The order of block was counterbalanced across participants.

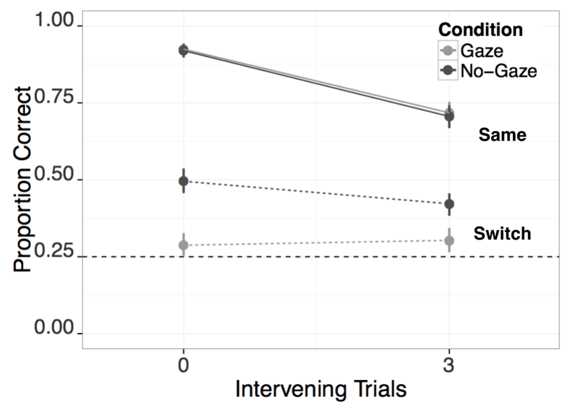

Figure 3: Accuracy on test trials in Experiment 2. Each datapoint represents 182 participants. Error bars indicate 95% confidence intervals computed by non-parametric bootstrap.

Results and Discussion

We followed the same analysis plan as in Experiment 1. First, we analyze performance on exposure trials to ensure that participants were using the referential cue and to test if the cue changed response times. Then we analyze performance on test trials to measure the effect of the presence of referential cues on the number of word-object links learners stored in memory.

Exposure trials Similar to Experiment 1, participants' responses on exposure trials differed from those expected by chance, exact binomial p(two-tailed) $< .001$, suggesting that eye gaze effective in directing attention to the target referent. Participants in Experiment 2 were numerically more consistent in their use of eye gaze with the live action stimuli in Experiment 2 compared to the schematic stimuli used in Experiment 1 (M1 = .76, M2 = .81). We also fit a linear mixed effects model to response times with the same specification as Experiment 1, finding main effects for gaze condition ($\beta = -1112.83$, $p < .001$) and interval ($\beta = -498.96$, $p < .001$) with faster responses in the gaze condition and in the the longer interval condition. The two-way interaction between gaze condition and interval was not significant, showing that gaze had the same effect on participants' response times at both intervals.

Test trials Figure 3 shows performance on test trials in Experiment 2. We replicate the main finding from Experiment 1: participants in the gaze condition performed worse on switch trials. We fit a mixed-effects logistic regression model[3] and found significant main effects of interval ($\beta = -0.55$, $p < .001$) and trial type ($\beta = -2.63$, $p < .001$). Participants were

[3] We fit models to both the unfiltered and filtered datasets and found no difference between the two analyses.

less accurate as the interval increased and on switch trials. In addition, the model showed significant two-way interactions between gaze condition and trial type ($\beta = -0.95$, $p < .001$) such that switch trials were more difficult after gaze exposure trials. Similar to Experiment 1, we did not find evidence of a boost to performance on same trials in the gaze condition.

Taken together, the data from Experiment 1 and 2 suggest that the presence of a referential cue reliably focuses learners' attention and shifts them towards single hypothesis tracking strategy. Changing to a live action stimulus set led to slightly higher rates of participants selecting the target of eye gaze on exposure trials, but did not result in a boost to performance on Same trials, providing additional evidence that the fidelity of participants' single hypothesis was unaffected in our paradigm by the presence of a referential cue.

Experiment 3

In Experiment 3, our goal was to move beyond manipulating the mere presence of a referential cue to a parametric manipulation of the strength of that cue. To accomplish this, we varied the reliability of eye gaze as a cue to reference. This design was inspired by experimental work showing that children are sensitive to the past reliability of those around them when deciding whom to ask for new information (Koenig, Clément, & Harris, 2004). By parametrically manipulating reliability, we hoped to test a clear prediction of our account: that learners will allocate attention and memory rationally in response to graded changes in the referential uncertainty present during learning.

Methods

Participants Participant recruitment, and inclusionary/exclusionary criteria were identical to those of Experiment 1 and 2 (excluded 4 HITs). 50 HITs were posted for each reliability level (0%, 25%, 50%, 75%, and 100%) for total of 250 paid HITs.

Design and Procedure Procedures were identical to those of Experiment 1 and 2. We modified our cross-situational learning paradigm to include a block of 16 familiarization trials (8 exposure and 8 test), which established the reliability of the referential cue. To establish reliability, we varied the proportion of same/switch trials that occurred during this familiarization block. Switch trials provide evidence that eye gaze is not a reliable predictor of the object that will appear at test. Participants either saw 0, 2, 4, 6, or 8 switch trials. After the familiarization block, participants completed a block of 16 trials (8 exposure and 8 test). Importantly, since we were no longer testing the effect of presence or absence of referential cues, all exposure trials in Experiment 3 included eye gaze, but this cue was more or less reliable depending on which familiarization block participants saw.

Results and Discussion

Exposure trials Participants reliably chose the referent that was the target of eye gaze at rates greater than those that would be predicted by chance p(two-tailed) $< .001$. We fit a mixed effects linear regression model and found a marginally significant effect of reliability level ($\beta = 1.03$, $p = .08$) such that as the reliability of the cue increased, participants were more likely to select the target of gaze on exposure trials, providing evidence that learners were sensitive to the reliability of the cue.

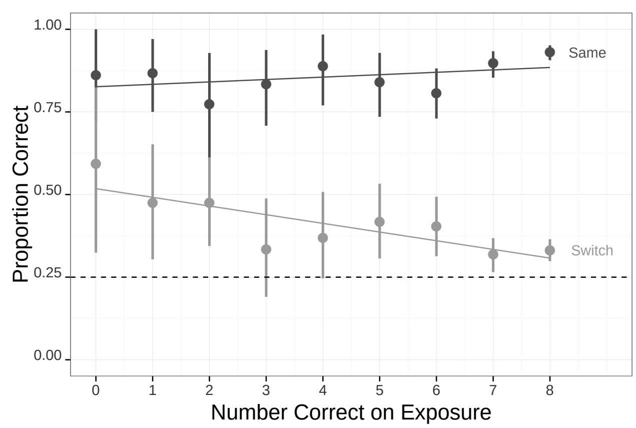

Figure 4: Accuracy on test trials in Experiment 3 for both trial types (Same and Switch) as a function of participants' use of eye gaze on exposure trials. Each datapoint represents approximately 50 participants. Error bars indicate 95% confidence intervals computed by non-parametric bootstrap.

Test trials Figure 4 shows participants accuracy on test trials within the test block. To quantify the effect of reliability on accuracy, we fit a mixed-effects logistic regression model and found a significant main effect of trial type ($\beta = -3.94$, $p < .001$), with participants responding less accurately on switch trials. In this analysis, we did not find a significant interaction between reliability and trial type.

Similar to Experiment 1, we would only expect to see an interaction between reliability and trial type if learners were using the gaze cue during exposure trials. Thus, we conducted a follow-up analysis where we modeled accuracy on test trials as a function of how often participants chose the target of eye gaze on exposure trials. We fit a mixed effects logistic regression model with the same specifications, but substituting accuracy on exposure trials for reliability condition as a predictor, and found a robust two-way interaction between performance on exposure trials and trial type ($\beta = -0.25$, $p < .001$) such that participants who were more likely to use the gaze cue performed worse on switch trials[4]. These analyses show that as a referential cue becomes more reliable, participants were more likely to use it, and that learners who used the referential cue were less likely to store multiple word-object links.

[4] In an earlier version of the experiment this analysis was post-hoc. So we ran a follow-up study, and all results reported here are from a planned analysis of that follow-up.

General Discussion

An ideal learner with unlimited attention and memory could track all possible word-object co-occurrences, making cross-situational word learning a simple problem of getting enough data points. But human learners are constrained by limited cognitive resources, making it important to decide which statistics to store from a learning moment. Recent work suggests that learners store a strong candidate hypothesis along with other possible word-object links with varying degrees of fidelity depending on the attention and memory demands present during learning (Yurovsky & Frank, under review).

In the current line of work, we extend these findings to show that an ecologically valid referential cue to word meaning—the speaker's eye gaze—focuses learners' attention and reduces the attention and memory allocated to other possible word-object links (Experiments 1 and 2). We also parametrically manipulated the reliability of the referential cue, and found that learners were more likely to use the cue as it became more reliable, and that when learners did use the cue, multiple hypothesis tracking decreased (Experiment 3). Interestingly, across all three experiments, reduced memory for alternative hypotheses did not result in a boost to performance on same trials. This pattern of data suggests that the presence of a referential cue selectively affected the number of word-object links stored in a given learning moment, but did not strengthen learners' memory for their candidate hypothesis.

There are several limitations to the current study that are worth noting. First, the social context we used was relatively impoverished. Here we isolated just a single cue to reference, eye gaze, using both a schematic and live action stimulus set. But real-world learning contexts are much more complex, providing learners access to multiple cues to reference such as eye gaze, pointing, and previous discourse. In fact, we did see a more reliable effect of referential cues when we used a live film, which included both eye gaze and head turn as opposed to the static, schematic stimuli. Second, we do not yet know how these results would generalize to young word learners. It is an interesting open question as to how children, who have even more limited cognitive resources, choose to allocate them during learning.

Our results fit well within a resource rational framework (Lieder et al., 2014), which attempts to push the rationality of computational-level models down to the psychological process level. In this framework, cognitive systems are thought to be adaptive in that they optimize the use of their limited resources, taking the cost of computation (e.g., opportunity cost of time or mental opportunity) into account. In the current work, learners showed evidence of adapting to the level of referential uncertainty in the learning context, changing how many word-object links they stored in memory.

Word learning proceeds despite the potential for high levels of referential uncertainty and learners' limited cognitive resources. Our work shows how referential cues can influence the allocation of cognitive resources, causing learners to store different numbers of word-object links from a labeling moment. Overall, these results increase our understanding of how social contexts support language acquisition.

Acknowledgments

We are grateful to the members of the Language and Cognition Lab for their feedback on this project. This work was supported by a National Science Foundation Graduate Research Fellowship to KM and an NIH NRSA Postdoctoral Fellowship to DY.

References

Baldwin, D. A. (1993). Infants' ability to consult the speaker for clues to word reference. *Journal of child language*, *20*(02), 395–418.

Clark, E. V. (2009). *First language acquisition*. Cambridge University Press.

Kanwisher, N., Woods, R. P., Iacoboni, M., & Mazziotta, J. C. (1997). A locus in human extrastriate cortex for visual shape analysis. *Journal of Cognitive Neuroscience*, *9*(1), 133–142.

Koenig, M. A., Clément, F., & Harris, P. L. (2004). Trust in testimony: Children's use of true and false statements. *Psychological Science*, *15*(10), 694–698.

Lieder, F., Goodman, N. D., & Griffiths, T. (2014). Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in Cognitive Science. forthcoming*.

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.

Quine, W. V. (1960). 0. word and object. *111e MIT Press*.

Smith, K., Smith, A. D., & Blythe, R. A. (2011). Cross-situational learning: An experimental study of word-learning mechanisms. *Cognitive Science*, *35*(3), 480–498.

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

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.

Vouloumanos, A. (2008). Fine-grained sensitivity to statistical information in adult word learning. *Cognition*, *107*(2), 729–742.

Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and done? optimal decisions from very few samples. *Cognitive science*, *38*(4), 599–637.

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

Yurovsky, D., & Frank, M. (under review). An integrative account of constraints on cross-situational learning.