Interactions between statistical aggregation and hypothesis testing mechanisms during word learning

Alexa R. Romberg and Chen Yu
{aromberg, chenyu}@indiana.edu
Department of Psychological and Brain Sciences, 1101 E. 10th Street
Bloomington, IN 47405 USA

Abstract

Adults, children and infants are all able to infer likely word meanings based on the relative frequency with which labels and referents appear together (e.g., Smith & Yu, 2007; Yu & Smith, 2008). However, the extent to which learners rely on aggregation of co-occurrence statistics vs. test specific hypotheses to infer mappings is currently a matter of significant uncertainty (Smith & Yu, 2012), exacerbated by the different experimental methods used to test learning mechanisms. Real world word learning is likely to involve a combination of statistical aggregation and active hypothesis testing. The current experiment investigates how these two learning mechanisms interact during word learning by having participants respond to a subset of items during a cross-situational word learning task. We find that hypothesis-testing is most effective when informed by statistical information and that the process of hypothesis-testing draws attention away from the remaining set of items.

Keywords: word learning; statistical learning; language acquisition; cross-situational learning; learning mechanisms

Introduction

In the past several years evidence has accumulated that infants, children and adults are able to aggregate information across multiple ambiguous contexts to disambiguate a word-referent mapping (e.g., Scott & Fisher, 2011; Smith & Yu, 2008; Suanda & Namy, 2012; Trueswell, Medina, Hafri & Gleitman, 2013; Yu & Smith, 2007). The learning mechanisms that learners use to do this *cross-situational* word learning are currently under debate, contrasting hypothesis testing (HT) with associative learning (AL; Koehne, Trueswell & Gleitman, 2013; Smith & Yu, 2012; Trueswell et al., 2013).

Within associative theories, statistical computations emerge from the strengthening and weakening of associations as a function of co-occurrence reliability and competition among associations. Within hypothesis testing theories, conceptually coherent hypotheses are confirmed or disconfirmed through a variety of procedures. While both of these frameworks are consistent with learners inferring mappings between words and referents that consistently co-occur, they offer fundamentally different characterizations of what it means to be a statistical learner. Accordingly, two quite different experimental paradigms have been used to investigate cross-situational word learning. In the paradigm supporting hypothesis testing, each trial has a One:Many structure, providing participants with a single auditory label and several (often 5) objects to view (Koehne et al., 2013; Trueswell et al., 2013). Versions in which participants have responded with a selection of the word's referent on each trial have produced similar results as versions in which they simply observe the trials (e.g., Trueswell et al., 2013).

The findings from the HT paradigm demonstrate clearly the power of correctly hypothesizing a word's meaning: Participants who select the correct referent for a label on a given trial are more likely to select the correct referent for that label on future trials than participants who made an incorrect selection (Koehne et al., 2013; Trueswell et al., 2013; see also Medina, Snedeker, Trueswell & Gleitman, 2011). Thus, researchers employing the HT paradigm have concluded that the participants learn by forming a single hypothesis for the likely label-object mapping and do not store multiple word-object occurrences at the same time. However, overall learning effects are relatively modest. Participants responded correctly on approximately 33% of the 12 trials in the final block of Experiment 1 (their 5th exposure to the label) in Trueswell et al., (2013), correctly selecting 1.6 more referents than expected by chance.

In the "AL" paradigm, each trial has a Many:Many structure. Participants are provided with multiple labels and objects, often 4 of each when adults are tested (e.g., Romberg & Yu, 2013; Suanda & Namy, 2012; Yu & Smith, 2007). The paradigm consists of a training phase, during which participants do not make any responses, and a test phase during which participants select the most likely referent from an array of objects presented in the study. Overall learning effects are higher in this paradigm compared with the HT paradigm. For example, adult participants responded correctly on 35% of the 18 test trials in Romberg & Yu (2013), correctly selecting 5.3 more referents than expected by chance (after 6 exposures to each label). The authors employing the AL paradigm have argued that participants learn by aggregating co-occurrence statistics across trials and encode multiple associations between labels and objects presented on individual trials.

The two paradigms set up very different processing tasks for the participants, as illustrated in Figure 1. The HT paradigm is characterized by repeated oscillation between processing and retrieval. The task structure either explicitly or implicitly encourages participants to engage in some computation (e.g., updating representations of statistical structure) and retrieval (i.e., bringing a hypothesis to explicit awareness) on every trial. The fact that only 1 label is provided means that participants have limited ability to "control" information flow because they have no choice about which label to select for attention and can only associate objects with 1 label on each trial. The AL paradigm, in contrast, is characterized by a continuous sequence of information during which time participants are unconstrained as to both which subset of items they select for attention or when or how they engage in computations on data collected. Participants may or may not engage in explicit hypothesis-testing in the AL paradigm, but they have the opportunity to encode associations between multiple objects and multiple labels, unlike the HT paradigm. Thus, one reason why researchers employing these different paradigms have reached different conclusions about the mechanisms that support learning is that the paradigms themselves vary in the extent to which they afford the different mechanisms.

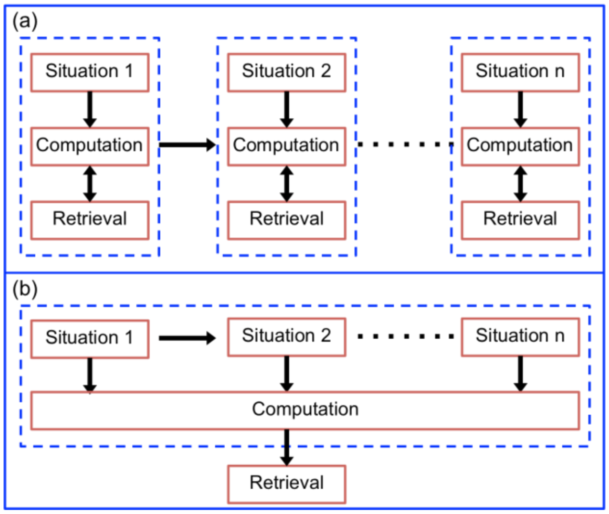

Figure 1. Schematic of the “Hypothesis Testing” (a) and the “Associative Learning” (b) experimental paradigms from a processing perspective.

Given that participants’ data encoding and computations are influenced by the task structure, we sought to create a new cross-situational word learning paradigm that afforded both hypothesis testing and statistical aggregation. A task that clearly affords both mechanisms is necessary to study how these mechanisms might influence one another. In addition, a hybrid paradigm is highly appealing as a more ecologically valid model of word learning. After all, it is not the case that leaners must actively guess a word’s meaning every time they hear it, or that they passively compute statistics without committing to particular mappings. Rather, real world learning involves both the implicit collection of statistical structure as well as explicit hypothesis testing.

A schematic of the hybrid (Mixed) paradigm from a processing perspective is provided in Figure 2. The task is structured so that One:Many trials are interleaved with Many:Many trials. This provides learners with opportunities to aggregate information across trials to potentially inform their hypotheses on the One:Many trials. Experiment 1 compares overall learning outcomes from the hybrid paradigm with those of the AL paradigm. Experiment 2 investigates the interaction between statistical aggregation and hypothesis testing within the hybrid paradigm.

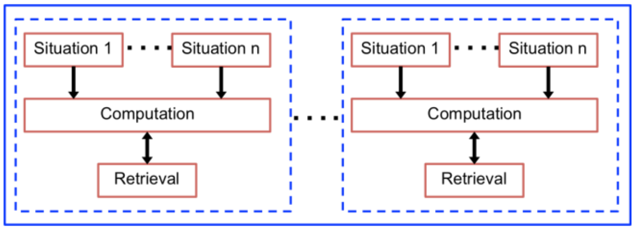

Figure 2. Schematic of the novel hybrid paradigm.

Experiment 1

The different learning rates from the HT and AL paradigms discussed in the introduction raises the question of whether the modest learning results reported by Trueswell et al. (2013) are because the One:Many trial structure reduces participants’ ability to efficiently use co-occurrence information available to them. If this is true, we would expect poorer performance in our hybrid paradigm than in a fully Many:Many paradigm. In Experiment 1 our goal was to test for differences between these paradigms while giving participants sufficient exposure that good performance should be attainable.

Method

Participants Thirty undergraduates participated for course credit (16 females).

Materials Auditory stimuli consisted of 36 nonce words synthesized with the Ivona voice Jennifer using the TextSpeaker program. Nonce words consisted of one or two syllables (264 ms to 795 ms in duration) and followed English phonotactics (e.g., *rud*, *vot*, *koom*, *vamey*, *genism*, *feddy*). Visual stimuli were 36 color photographs or 3D models of novel or real objects that were not readily nameable. Images, approximately 3” square, were displayed on a white background in the corners of a 17” monitor.

Experiment Design Each participant completed 2 cross-situational word learning tasks within the 40 minute session. Each task included 18 different label-object pairs and consisted of a training phase, whose structure varied between tasks, and a test phase, whose structure was the same across tasks. Examples of the training phases are given in Figure 3.

The *Mixed Condition* had 2 identical blocks of 35 training trials. Within each block, there were 27 “4x4” (i.e., Many: Many) trials on which 4 objects were shown and 4 labels were played. Participants were told that the order of the labels did not correspond in any way with the positions of the objects on the screen. Eight trials within each block were “1x4” trials (i.e., One:Many) on which 4 objects were shown and 1 label was played. On 1x4 trials only, the word “Select” appeared in the middle of the screen, instructing participants to select the object they thought was the most likely referent of the word. No feedback was given on their selection. The 1x4 trials were interleaved with the 4x4 trials, with approximately 3 4x4 trials between each 1x4 trial. A different label was played on every 1x4 trial, so that 8 labels in total were "probed". For all training trials, the referent for the label(s) played was always present.

This design means that the 8 label-object pairs that were probed in the 1x4 trials were presented 14 times during the training, 12 times on 4x4 trials and twice on 1x4 trials. The 10 unprobed label-object pairs (i.e., whose labels were not played on 1x4 trials) were presented on 12 4x4 training trials. All 18 objects were distributed as evenly as possible as foils across the 1x4 trials: Within each block, all 8 of the probed objects appeared as a foil on another 1x4 trial, 6 of the unprobed objects were foils on two 1x4 trials and 4 unprobed objects were foils on a single 1x4 trial.

The *Many-Only Condition* had two identical blocks of 29 training trials. All trials were 4x4 trials and no responses were collected during training. These two within-subjects conditions were designed to be as parallel as possible. Each label and object within each condition was randomly assigned to a number from 1 to 18. These 18 items were pseudo-randomly grouped into the 27 4x4 training trials with the constraint that no more than 1 pair was presented on consecutive trials. Thus, while the actual objects and labels were unique to each task, the same sequence of 27 4x4 trials were used across both conditions (e.g., pair 1 was presented on the same 4x4 trials and the label and object were presented in the same positions within each of those trials for both tasks). To equalize the number of repetitions for each pair across conditions, the 8 pairs in the Many-Only condition that corresponded to the Probed items in the Mixed condition were presented on one additional 4x4 trial within each block. These two extra 4x4 trials were inserted 1/3 and 2/3 of the way through the 27 other trials.

Training was followed immediately by test for both conditions. On each test trial, 1 label was played and participants selected its most likely referent from an array of all 18 objects presented in that condition. Each label was tested once with the order randomized for each participant.

To control for item effects each participant was randomly assigned to one of 2 different stimulus sets. The stimulus sets contained the same 36 label-object pairs, but the pairs were distributed between the conditions in different random assignments for each set. Additionally, participants were randomly assigned to one of 2 different trial orders. The trial orders varied both in the order of items within and across trials and in the exact placement of the 1x4 trials.

Procedure Participants were tested individually. They were given an overview of the experiment and informed consent was obtained. The order of the conditions was randomized for each participant. Each word-learning task was preceded by a set of slides with the specific instructions for that task. All participants completed all conditions.

Results and Discussion

Accuracy in both conditions was close to ceiling, with 1/3 of the participants getting 17 or 18 of the 18 mappings correct for each task. There was no difference between conditions in proportion of mappings learned (Response: *M*=0.861, *SD*=0.29; No-Response: *M*=0.889, *SD*=0.35). The number of training trials in the current experiment was approximately double that used in previously reported studies using the Many:Many paradigm (e.g., Yu & Smith, 2007; Romberg & Yu, 2013). The performance is a marked increase over the mean of 35% reported by Romberg & Yu (2013), even though all training trials were highly ambiguous.

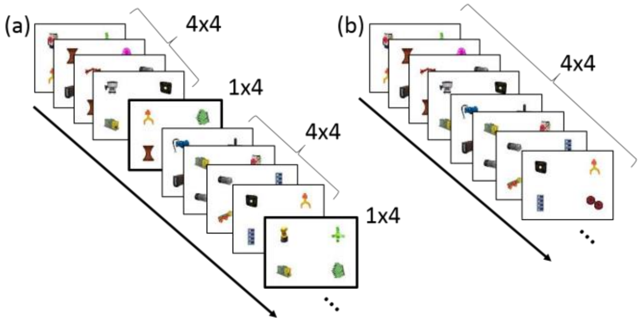

Figure 3 Schematic of the sequence of training trials used in Experiments 1 & 2 in the (a) Mixed and Mixed-No Response conditions and (b) Many-Only condition. "4x4" trials presented 4 words and 4 objects and "1x4" trials presented 1 word and 4 objects.

These results demonstrate that statistical learning can be a highly efficient process. The training phase lasted less than 15 minutes, yet many participants were able to learn 18 new mappings. Also important is that performance across the two conditions was very similar, demonstrating that rapid learning is possible from both Many:Many trials alone and from a mixture of One:Many and Many:Many.

Experiment 2

The ceiling level performance in Experiment 1 prevented investigation into how associative learning and hypothesis-testing mechanisms might interact. Experiment 2 addresses this question by reducing the number of training trials so that any potential differences would be detectable, either between the conditions or between the different types of items within each condition. We also include a third condition, *Mixed-NoResponse*, with the same structure as the Mixed condition but in which participants are not asked to actually select a referent on the 1x4 Probe trials. This condition addresses the possibility that the act of responding in itself may drive any observed differences between the Mixed and the Many-Only conditions.

A second way to test for interactions between the two learning mechanisms is to compare learning within and between conditions on the Probed items (for which participants are encouraged to form hypotheses in the two Mixed conditions) and the Unprobed items (for which no hypothesis-testing was encouraged in any condition). If hypothesis-testing is the primary mechanism that participants are using regardless of trial structure, then performance on the Probed and Unprobed items should be equivalent within the Mixed conditions. However, if participants rely on both hypothesis-testing and associative learning there may be a difference in learning outcome across the two item types within the Mixed conditions.

Method

Participants Eighty-seven undergraduates participated for course credit. Three participants were excluded for scoring less than 6% correct in each of the three conditions. Data from 4 participants was lost due to technical error. The final sample consisted of 80 participants (42 females).

Materials The stimuli included the 36 words and objects used in Experiment 1 as well as 18 additional words and objects created in the same manner.

Experimental Design and Procedure Participants each completed 3 cross-situational word learning tasks within the 45 minute session. The Mixed and Many-Only conditions were identical to those in Experiment 1, but participants were only given 1 block of training trials for a total of 35 and 29 training trials, respectively. The third condition, *Mixed-NoResponse* was identical to the Mixed condition except that participants were not instructed to make a selection on the 1x4 trials. The procedure was the same as Experiment 1.

Results and Discussion

Hypothesis-testing influences associative learning

In contrast to Experiment 1, participants learned more mappings in the Many-Only condition than those that encouraged hypothesis-testing. Participants learned the highest proportion of mappings in the Many-Only condition (M=0.456, SD=0.312) and lowest in the Mixed condition (M=0.394, SD=0.300). The Mixed-NoResponse condition fell between the other two (M=0.403, SD=0.294). A logistic mixed-effect model was fit to the raw data with Condition as a fixed effect and random effects of Subject on the intercept and on the slope of Condition. Condition was dummy coded with Many-Only as the reference. The model confirmed a significant contrast between the Mixed and Many-Only conditions (b=-0.428, z=2.34, p=0.02) and a marginally significant contrast between the Mixed-NoResponse and Many-Only conditions (b=-0.336, z=1.84, p=0.07).

Within the Mixed conditions, however, hypothesis-testing was related to better learning. The mean proportion of items correct for the Probed and Unprobed items for each condition is provided in Figure 4. Note that in the Many-Only condition the Probed items were not actually probed (i.e. labels presented on 1x4 trials) but were presented one extra time relative to the Unprobed items, as in the other conditions. Thus, the advantage for Probed items in the Mixed conditions cannot be due to the extra exposure.

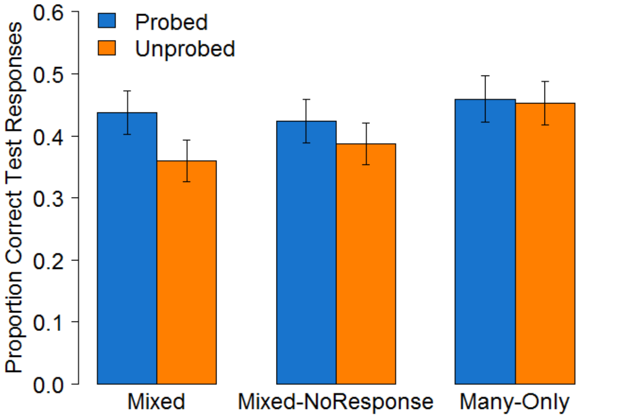

Figure 4. Mean proportion of correct test responses (SE) for each condition and item type in Experiment 2.

Accuracy on Probed items is equivalent across the 3 conditions while accuracy on Unprobed items is lower in the Mixed conditions. A logistic mixed-effect model confirms this result. The model had Item Type (Probed vs. Unprobed) and Condition (Mixed, Mixed-NoResponse and Many-Only) as fixed effects and random effects of Subject on the intercept and on the Item Type X Condition interaction. A significant contrast between the Many-Only and Mixed conditions (b=-0.654, z=3.56, p=0.004) was qualified by a significant interaction between that contrast and Item Type (b=0.534, z=2.02, p=0.043). The contrast between Many-Only and Mixed-NoResponse was also significant (b=-0.410, z=2.26, p=0.024), but the interaction between that contrast and Item Type was not (b=0.187, z<1). Models fit to the individual conditions confirm a significant difference between item types for the Mixed condition (b=0.481, z=3.70, p<0.001), a marginal difference for the Mixed-NoResponse condition (b=0.219, z=1.71, p=0.09) and no difference in the Many-Only condition (b=0.043, z<1).[1]

The decrement in overall learning, and specifically in learning of the Unprobed mappings, suggests that encouraging hypothesis-testing on a subset of items negatively influenced learning of the rest of the set. This implies that the processes participants used on the 1x4 trials were different than those used on the 4x4 trials. The fact that the Mixed-NoResponse condition was consistently in between the other two suggests that being asked to respond exacerbated effects of the 1x4 trials, perhaps by making hypotheses more explicit or by limiting other processing that participants could do.

Associative learning influences hypothesis-testing

To examine the interactions between associative learning and hypothesis-testing further, the Mixed condition was

[1] There are fewer Probed items (8) than Unprobed items (10), raising the concern that the lower accuracy for Unprobed items is driven by the smaller gain for each correct item. To address this, a series of logistic mixed effect models were fit to responses for the 8 Probed items and each of the 45 unique subsets of 8 Unprobed items for the Response-Mixed and Many-Only conditions. P-values for the interaction were < 0.05 in 33 of the 45 models. All 45 models fit to each the individual conditions found a significant difference between Probed and Unprobed items for the Mixed condition and no difference for the Many-Only condition.

analyzed in detail. Prior findings suggest that forming a correct hypothesis can be an important step in word learning (Medina et al., 2011; Trueswell et al., 2013; Yu, Zhong & Fricker, 2012). Consistent with this, in the Mixed condition participants 'test accuracy on the 8 Probed items was influenced by whether they had selected the correct object on the corresponding 1x4 trial during training. Overall, participants selected the correct object on half the 1x4 trials during training (*M*=0.516, *SD*=0.229). Participants were significantly more accurate on the test for Probed items they had gotten correct during training (*M*=0.550, *SD*=0.389) than on Probed items they had gotten incorrect (*M*=0.333, *SD*=0.367). A logistic mixed-effects regression model confirms this difference (*b*=1.54, *z*=7.47, *p* < 0.001).

The test mean for Probed items that participants got incorrect during training is very close to the test mean for the Unprobed items in this condition and well above the chance baseline of 0.056 (random selection from the 18 items present at test). This result demonstrates that incorrect hypotheses are not as fatal as prior studies have suggested (Medina et al., 2011). Rather, participants draw on the co-occurrence structure to infer likely mappings regardless of whether they were encouraged to form a hypothesis for that item.

If the hypotheses learners made on 1x4 trials were informed by the co-occurrence statistics of the items present on the trial, two effects should be present. First, the number of prior exposures to the probed label-object pair should positively predict accurate selection, since learners would have accrued more examples of the pairing. Second, the number of times the probed pair had previously co-occurred with each of the other objects present on the trial should negatively predict accurate selection, since those objects have a partial association with the probed label. Both of these effects were found in the current data. A logistic mixed-effects model was fit to participants' accuracy on the 8 1x4 training trials. Probe Exposure (the number of repetitions of the probed label-object pair up to that point in the experiment) and Total Foil Co-occurrence (the sum of the number of times each of the 3 foil objects had co-occurred with the probed item up to that point in the experiment) were entered as fixed effects and random effects of Subject and Order on the intercept were included (Order was included because the values of Probe Exposure and Foil-Co-occurrence vary between the two different trial orders used). As predicted, Probe exposure had a significant positive effect on accuracy (b=0.486, *z*=5.56, *p*< 0.001) and Total Foil Co-occurrence had a significant negative effect (*b*=-0.256, *z*=2.14, *p*= 0.032).

General Discussion

Interactions between learning mechanisms

In Experiment 1, participants performed close to ceiling on both the Mixed and the Many-Only conditions. However, decreasing the amount of training in Experiment 2 resulted in significant differences between the conditions. This pattern suggests that encouraging hypothesis-testing during statistical word learning initially slows learning but that this disadvantage disappears with sufficient training.

The initial decrement in learning rate may be due to the process of hypothesis-testing interfering with participants' more implicit associative learning. Participants learned a smaller proportion of the Unprobed items in the Mixed condition than in the Many-Only condition, indicating that forming explicit hypotheses about some items influenced how effectively participants tracked co-occurrences for the *other* items. When participants were not required to actually click on a referent on the One:Many trials, the negative effects of hypothesis testing were partially ameliorated. The specific reason for this effect of clicking depends on the pathway for the influence of hypothesis-testing on associative learning (considered below). However, the most general explanation is that because participants had no *specific* task on the One:Many trials when a response was not required, the processes they used on the One:Many trials were more similar to those used on the Many:Many trials.

Why might hypothesis-testing be detrimental to associative learning? One possibility is that having participants make a response to a subset of items caused them to focus their attention primarily on that subset. Less attention to the other items would result in learning less about them. While intuitively appealing, this explanation is inconsistent with the current data. Participants did not know ahead of time which labels would be presented on the One:Many trials. This account would therefore make two predictions: First, that participants best learn the items that were probed first, since participants would have the most exposures to those items post-probe. Second, that accuracy in the test phase would be independent of response on the probe trials, since the primary contribution of the probe trial would be to spur participants to attend to the probed label. However, both of these predictions are opposite what was found. Accuracy at test depended strongly on the accuracy of participants' responses on the One:Many trials. Participants were more likely to respond correctly on later-probed items, suggesting that much of the learning occurred *before* the actual probe trial, rather than after.

A second possible explanation is that the *process* of explicit hypothesis testing interferes with information aggregation in some way. As each trial unfolds, participants encode (at least some) of the labels and objects presented and presumably do computations to update the associative strength between various label-object pairs. Such updating likely involves not just the items present on the current trial, but extends to possible inferences about other items (Romberg & Yu, 2013). When no response is required, participants can spend the entire trial duration in these processes (see Figure 1b). In such cases participants receive a fairly uniform flow of information. Requiring a response punctuates this information processing with additional retrieval and response (motor) processes (see Figure 1a).

The current data cannot tell us exactly *how* retrieval and response interfere with information processing. Many non-mutually exclusive possibilities exist. Retrieval and response may interfere with encoding and updating directly by interrupting the other processes. Retrieval of incorrect information may cause incorrect updating (e.g., strengthening an unattested association or one that the current evidence should weaken). Such incorrect updating may be one reason that forming an incorrect hypothesis had such a negative effect on learning in both the present study and prior work (Koehne et al., 2013; Medina et al., 2011; Trueswell et al., 2013). It is also possible that focusing attention on a response causes forgetting of previous associations. Additional experiments are required to explore these possibilities.

Internally directed attention facilitates learning

While it seems reasonable to infer likely learning mechanisms, such as hypothesis-testing and associative learning from the affordances of the different paradigms, our current data cannot ultimately tell us what processes learners employed. However, the structure of each paradigm and the instructions participants received are completely known and the data reveal clear differences between paradigms.

Our results suggest that learning is facilitated when participants freely choose how to allocate their attention and computational processing. While the Mixed condition was designed to encourage explicit hypothesis testing for a subset of items, participants may very well have formed and tested hypotheses in the Many-Only condition as well. The primary difference between the conditions was the extent to which participants 1) controlled the distribution of their attention between the items and 2) selected *which* items to form explicit hypotheses about.

On Many:Many trials, participants could decide whether to attempt to retrieve specific labels for objects during the trial or to make an explicit link between a particular label and object. They could choose to select some labels and objects for attention while ignoring others. Or they could attempt to store as much information from each trial as possible. In contrast, on One:Many trials participants had to focus their attention on the one label provided and (when a response was required) attempt to match the label with one of the objects present.

There is strong evidence from Experiment 2 that participants' hypotheses are informed by co-occurrence statistics. The lack of explicit structure on Many:Many trials, as well as the large amount of information available, allows learners to (consciously or unconsciously) flexibly adapt their attention to particular information levels. For example, they may neglect labels/objects that are already known or "out of reach" and attend largely to those for which they have some partial knowledge. There is evidence from other domains that attention is biased toward input with a moderate rate of information (e.g., Kidd et al., 2012).

The faster learning rate does not necessarily make the AL paradigm a better experimental model for word learning. Indeed, the Mixed condition was specifically designed to more closely model the type of information gathering punctuated by retrieval that is involved in word learning in the real world (as well as learning in other domains). The present results illustrate how incredibly sensitive learners are to the statistical structure of their environment. Understanding how hypotheses influence and are informed by statistical computations is a critical step for advancing the science of learning.

Acknowledgments

This work was supported by NIH R01 HD056029 to C. Yu and an NIH Ruth Kirschstein NRSA training grant to Indiana University: T32 HD007475-17. The authors thank Morgan Doelling and Drake Belt for assisting with data collection.

References

Kidd, C., Piantadosi, S. T., & Aslin, R. N. (2012). The Goldilocks effect: human infants allocate attention to visual sequences that are neither too simpler nor too complex. PLoSONE 7:e36399.doi:10.1371/journal.pone. 0036399

Koehne, J., Trueswell, J. C., & Gleitman, L. R. (2013). Multiple proposal memory in observational word learning. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 25th Annual Conference of the Cognitive Science Society* (pp.805-810). Austin, TX: Cognitive Science Society.

Medina, T. N., Snedeker, J., Trueswell, J. C., & Gleitman, L. R. (2011). How words can and cannot be learned by observation. *PNAS*, *108*(22), 9014-9019.

Scott, R. M., & Fisher, C. (2011). 2.5-Year-olds use cross-situational consistency to learn verbs under referential uncertainty. *Cognition*, *122*, 163-180.

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

Suanda, S. H., & Namy, L. L. (2012). Detailed Behavioral Analysis as a Window Into Cross- Situational Word Learning. *Cognitive Science*, *36, 545-559.*

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*, 126-156.

Romberg, A. R. & Yu, C. (2013). Integration and inference: Cross-situational word learning involves more than simple co-occurrences. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 25th Annual Conference of the Cognitive Science Society* (pp.1235-1240). Austin, TX: Cognitive Science Society.

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

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

Yu, C., Zhong, Y., & Fricker, D. (2012). Selective attention in cross-situational statistical learning: evidence from eye tracking. *Frontiers in Psychology*, 3:148 doi: 10.3389/fpsyg.2012.00148