BUCLD 39 Proceedings
To be published in 2015 by Cascadilla Press
Rights forms signed by all authors

Statistical Aggregation and Hypothesis Testing Mechanisms Interact during Word Learning

Alexa R. Romberg and Chen Yu

1. Introduction

Referential utterances are by their nature ambiguous to novice language learners. Each utterance consists of multiple layers of information that must be decoded: 1) the linguistic structure (how the sounds and words should be packaged into meaningful units), 2) the world structure (how people, objects and actions in the world relate to one another and which is the current focus of attention) and 3) the relation *between* the language and the world.

Tracking structure over time, frequently referred to as "statistical learning," is one way to resolve these ambiguities and gain access to additional patterns within and across domains. For example, in statistical learning of word boundaries, human learners, including infants, track the relative predictability of sequences of syllables in the language stream. Such tracking facilitates word segmentation because individual words can be used in many difference linguistic contexts. Thus, syllable sequences within words will tend to travel together, or be more predictable, than the sequences that bridge word boundaries (e.g., the word *sequence* was preceded by the words *of*, *syllable* and *the* in the sentences above, making the syllable bigram *se-quence* more predictable than *of-se*, *the-se* or *ble-se*).

Temporal cues in the auditory domain, such as the relative predictability of syllable sequences discussed above, may be augmented by additional cues from other domains or modalities. For example, there is evidence that word segmentation and word learning (mapping segmented words to visual referents) are more efficiently achieved when the two processes are done simultaneously rather than sequentially (Yurovsky, Yu & Smith, 2012). Such learning, of tracking language and world structures over time, has been termed "cross-situational" learning, with the different world-situations, or contexts, serving the disambiguating role that the different syllable contexts serve in statistical word segmentation.

---

* A. Romberg, University of Maryland, aromberg@umd.edu; C. Yu, Indiana University. A previous version of this work was presented at the Cognitive Science Society Meeting in July, 2014. 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.

Cross-situational learning of object labels has recently been an intense focus of attention amongst researchers interested in word learning. In the process of cross-situational word learning, learners link auditory words and visual referents that predictably co-occur together. There is a growing body of literature demonstrating that infants (e.g., Smith & Yu, 2008; Vlach & Johnson, 2013) children (e.g., Scott & Fisher, 2011) and adults (e.g., Romberg & Yu, 2013; Suanda & Namy, 2012; Trueswell, Medina, Hafri & Gleitman, 2013; Yu & Smith, 2007) are capable of mapping auditory labels to visual objects or actions by relying on cross-situational information that disambiguates word meanings across multiple exposures.

The cross-situational word learning literature stands in contrast to, though not necessarily in opposition to, the classic literature on fast-mapping of object labels. In fast mapping, learners draw on their prior knowledge and the specific structure of the referential event in order to form a hypothesis about a word's referent from a single instance. (e.g., Markman, 1990; Horst & Samuelson, 2008) Such mappings could be thought of as explicit hypotheses that learners can test when they hear the word at later times or by producing the word themselves. Thus, while the cross-situational word learning literature has emphasized the aggregation of information over time, the fast mapping literature has emphasized the formation of testable hypotheses from a single instance.

However, the specific learning mechanisms that contribute to cross-situational word learning have been a matter of significant debate. Some authors have argued that cross-situational word learning is a specific case of general associative or statistical learning, as detailed above (e.g., Romberg & Yu, 2013; Smith & Yu, 2012; Yurovsky, Fricker, Yu, & Smith, 2014; Yu & Smith, 2007). In this framework, information is aggregated over time through associative mechanisms that are likely implicit, at least at their beginning stages. Words are associated with objects with which they co-occur using the same mechanisms by which words are associated with other words, and objects associated with other objects. A learner acquires a meaning for a word when the association between that word and a referent is relatively strong compared with associations between that word and other referents, and that referent and other words.

Other authors have argued that cross-situational word learning more closely aligns with fast-mapping mechanisms, in which learners form a single hypothesis for a label-object mapping and confirm or reject that hypothesis with additional information (e.g., Koehne, Trueswell & Gleitman, 2013; Medina, Snedeker, Trueswell & Gleitman, 2011; Trueswell, Medina, Hafri & Gleitman, 2013). Under this hypothesis-testing framework, general co-occurrence or statistical information about label-object co-occurrences is neglected and does not inform hypothesis formation or testing. Rather, hypotheses are formed randomly and confirmed or rejected in a binary fashion based on whether the hypothesized referent is present when the word is heard (Trueswell et al., 2013).

The associative learning and the hypothesis-testing frameworks offer radically different characterizations of the learning process, the information available for inference and the types of representations employed by learners.

However, both frameworks predict that the outcome of learning should be linking words and referents that consistently co-occur, making the accounts challenging to distinguish from behavioral learning outcomes alone (Smith & Yu, 2012). Further, we suggest that it is unlikely that *only* associative learning or *only* hypothesis-testing mechanisms are employed by all learners at all times for cross-situational word learning. Such a claim requires ignoring one of the two large literatures discussed above: that of statistical learning of language and world structures or that of rapid "fast-mapping" inference of word meanings from a single instance.

Rather than attempt to adjudicate between these two learning frameworks, the current experiment was designed to investigate potential interactions between explicit hypothesis-testing and implicit associative learning mechanisms. To this end, we employed a novel experimental cross-situational word learning paradigm that included both opportunities for participants to freely aggregate information and for them to test hypotheses about a particular label's referent.

1.1 Study Overview

The present study was designed to address 3 questions about possible interactions between hypothesis testing and statistical learning mechanisms, each of which has been raised in the prior literature: 1) *Whether explicit hypothesis testing interferes with statistical learning or whether confirming hypotheses boosts performance.* To the extent that statistical learning may be an implicit process, explicit strategizing and hypothesizing may interfere with it. This is one reason why authors sometimes employ cover tasks during statistical sequence learning studies (e.g., Arciuli & Simpson, 2012; Saffran, 2002). However, there is also a literature suggesting that testing knowledge during learning boosts later retrieval over further studying (e.g., Karpicke & Roediger, 2006). The cross-situational word learning task is generally set up with explicit instructions (i.e., participants know that they are trying to link the objects and words) but without any direct instruction (i.e., all information is ambiguous and there is likely more information on any particular trial than can be explicitly encoded and maintained). Thus, it's not obvious from the task structure whether explicit hypothesizing might be distracting, as is suspected during implicit learning or might serve to consolidate learned information, as in explicit studying.

2) *Whether incorrect hypotheses interfere with learning.* In a recent study investigating cross-situational word learning, Medina and colleagues proposed that once an incorrect hypothesis about a word's referent was formed, it was very difficult for learners to recover and form a correct hypothesis (Medina et al., 2011). The authors suggested that this led to complete failure to learn unless sufficient time (24 hours) passed between the initial incorrect hypothesis and future information. The current study provides an opportunity to test whether this finding holds using a different cross-situational test procedure.

3) *Whether hypotheses are informed by statistical structure or are isolated from co-occurrence information.* The idea that hypotheses are isolated from statistical information is central to the Propose But Verify model of cross-situational word learning outlined by Trueswell and colleagues (Trueswell et al., 2013). The claim of isolation from other information is a primary point of tension between that model and statistical/associative learning accounts of cross-situational word learning. If hypotheses are found to be informed by statistical structure, then it seems likely that these two theories of cross-situational word learning may be more closely aligned than has been depicted in the literature thus far.

Participants each completed three different versions of a cross-situational word learning procedure. One was modeled off of a widely-reported paradigm that included no instructions for hypothesis-testing. The other two versions were similar to the first but also contained trials that we expected to encourage explicit hypothesis testing (for a subset of items).

2. Method

2.1 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).

2.2 Materials

Auditory stimuli consisted of 54 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 54 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.

2.3 Experiment Design

Each participant completed 3 cross-situational word learning tasks within the 45 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. A schematic of the training trial structure is provided in Figure 1.

The *Mixed-Response Condition* had 35 training trials. Twenty-seven of these trials were "4x4" 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. Participants did not have

a specific task during the 4x4 trials other than to attempt to learn which label went with which object. Eight trials were "1x4" hypothesis-testing trials, 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.

The *Mixed-No Response* condition was identical to the Mixed-Response condition, except that participants were not asked to respond on the 1x4 hypothesis-testing trials. The word "select" did not appear in the middle of the screen for these trials in this condition and participants' only instruction was to attempt to learn which words referred to which objects.

This design means that in the two Mixed condition, the 8 label-object pairs that were probed in the 1x4 trials were presented 7 times during the training, 6 times on 4x4 trials and once on 1x4 trials. The 10 unprobed label-object pairs (i.e., whose labels were not played on 1x4 trials) were presented on 6 4x4 training trials. All 18 objects were distributed as evenly as possible as foils across the 1x4 trials: 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 *4x4-Only Condition* had 29 training trials. All trials were 4x4 trials and no responses were collected during training. This condition was designed to align closely with prior studies of cross-situational word learning (e.g., Romberg & Yu, 2013; Yu & Smith, 2007; Yu, Zhong & Fricker, 2012). While learners may have decided on their own to test hypotheses during training, they were not explicitly encouraged to do so in this condition.

These three 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 condition, 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 4x4-only condition that corresponded to the Probed items in the Mixed conditions 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 all 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.

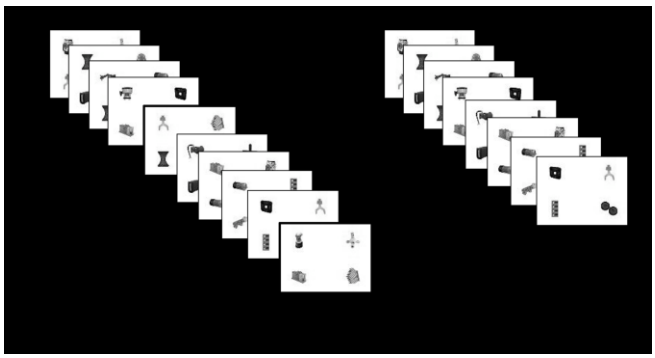

Figure 1. Schematic of the sequence of training trials used in the (a) Mixed and Mixed-No Response conditions and (b) 4x4-Only condition. "4x4" trials presented 4 words and 4 objects and "1x4" trials presented 1 word and 4 objects.

2.4 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 version of the word-learning task was preceded by a set of slides with the specific instructions for that task. All participants completed all conditions.

2.5 Predictions

To test for the influence of explicit hypothesis-testing trials on cross-situational word learning, we compared test accuracy across the three within-subjects conditions. If hypothesis-testing is beneficial to learning, as in the testing-restudy framework, we would expect 1) better word learning in the Mixed-Response condition than the 4x4-only condition and 2) better learning of probed words than unprobed words within the Mixed-Response but no the 4x4-only condition. However, the if hypothesis-testing is detrimental to cross-situational word learning, we would expect the opposite pattern, with poorer performance on items and conditions that include the hypothesis-testing trials.

To address the specific question of how disruptive incorrect hypotheses are to learning, we investigated the relation between accuracy on probe trials during training and accuracy at test within the Mixed-Response condition. If incorrect

hypotheses are exceptionally difficult to recover from, we would expect that those probed words to which participants responded incorrectly during training would not be learned at test.

Finally to address whether participants' hypotheses are informed by the co-occurrence structure of the labels and objects, we fit a logistic regression model to participants' accuracy on probe trials during training in the Mixed-Response condition. If hypotheses are formed randomly, as suggested by the Propose-But-Verify model, we should not be able to predict correct responses from co-occurrence information. If hypotheses are informed by statistical structure, then we would expect two factors to be particularly useful for predicting accurate responding. The first factor is the number of prior co-occurrences of the target word and object, as this represents the strength of the "correct" association between the word and its referent. The second factor is the number of co-occurrences between the target word and the 3 foil objects that are present on the 1x4 trial. This represents the strength of the "spurious" associations between the target word and the other objects. Stronger "correct" and weaker "spurious" associations should predict better chance of choosing the correct referent on the hypothesis-testing trial.

The Mixed-NoResponse condition was included as a measure of how important the actual responding is to any overall patterns that we find in the number of words learned. If the physical response does not particularly matter, we would expect the two Mixed conditions to pattern together. If, however, the physical response does influence the learning process, the two Mixed conditions may show different patterns of learning.

3. Results

3.1 Multiple effects of hypothesis testing on learning

Participants learned more mappings in the 4x4-Only condition (M=0.456, SD=0.312) than the Mixed-Response condition (M=0.394, SD=0.300), as measured by proportion of correct responses on the 18-AFC test. The Mixed-NoResponse condition fell between the other two (M=0.403, SD=0.294).[1] *Within* both the Mixed-Response and Mixed-NoResponse conditions, however, hypothesis-testing was related to better learning. Labels that participants heard on both 1x4 and 4x4 trials (Probed items) were better learned than labels heard only on 4x4 trials (Unprobed items). The mean proportion of correct 18-AFC test responses for the Probed and Unprobed items for each condition is provided in Figure 2. Note that in the 4x4-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

[1] A second study reveals that this interference is a somewhat transient effect (Romberg & Yu, 2014, Experiment 1). When the number of training trials is doubled, overall learning is equivalent in the 4x4-Only and Mixed-Response conditions. With longer exposure, participant accuracy was close to ceiling in both conditions.

advantage for Probed items in the Mixed conditions cannot be due to the extra exposure.

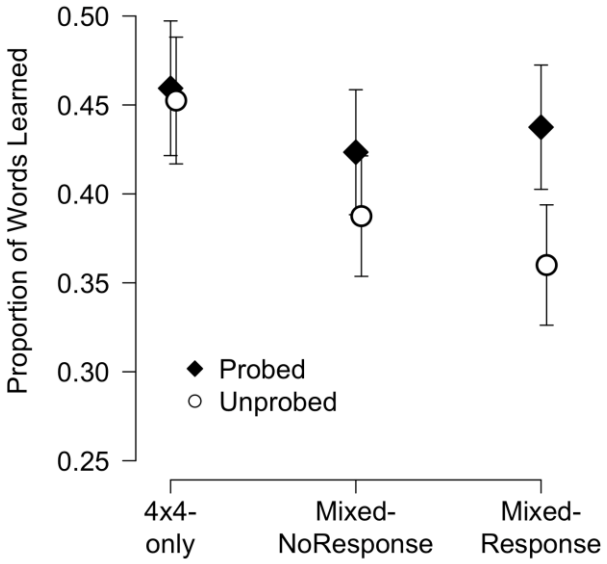

Figure 2. Mean proportion of correct test responses (SE) for each condition and item type.

Accuracy on Probed items was equivalent across the 3 conditions while accuracy on Unprobed items was lower in the Mixed conditions. A logistic mixed-effect model confirms this result. The model had Item Type (Probed vs. Unprobed) and Condition (Mixed-Response, Mixed-NoResponse and 4x4-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 4x4-Only and Mixed-Response 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 4x4-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-Response 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 4x4-Only condition (b=0.043, z<1).[2]

[2] 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

As noted above, previous research has found the effects of forming a correct or incorrect hypothesis during word learning to be quite divergent. Several prior findings suggest that forming a correct hypothesis can be an important step in word learning, whether that hypothesis is inferred from referent selection, as in the current study, or by eye gaze measures (Medina et al., 2011; Trueswell et al., 2013; Yu, Zhong & Fricker, 2012). However, some authors have posited that incorrect hypotheses essentially derail learning (Medina et al., 2011).

To investigate how hypothesis accuracy was related to learning, we broke down participants' 18-AFC test performance for the Probed items based on whether they had responded correctly to that item on its 1x4 trial during training in the Mixed-Response condition. 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, $p < 0.001$).

Our data reveal a powerful effect of hypothesis accuracy on word learning. Items for which participants responded correctly during training were much more likely to be retained at test than those to which participants' responded incorrectly. However, we did not find evidence that incorrect hypotheses are exceptionally difficult for learners to recover from. The test mean for Probed items that participants got incorrect during training was very close to the test mean for the Unprobed items in this condition (see Figure 1) and well above the chance baseline of 0.056 (random selection from the 18 items present at test).

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 is consistent with the idea that explicit hypothesis-testing disrupts the aggregation of co-occurrence statistics. However, when hypotheses were correct, participants were more likely to retain those mappings at test, consistent with the increased retention found from testing.

3.2 Hypotheses are influenced by co-occurrence statistics

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

---

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 4x4-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 4x4-Only condition.

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).

4. Discussion

4.1 Hypothesis-testing and retrieval effects in cross-situational word learning

We entered this investigation with two competing frameworks for how explicit hypothesis-testing might interact with statistical learning in cross-situational word learning. The first framework was that the explicit processes involved in hypothesis-testing would interfere with the, believed largely implicit, aggregation of statistical information. The second was that hypothesis-testing would provide an opportunity to "test" learners' knowledge of label-object mappings and that such testing should be beneficial to learning overall. Interestingly, the pattern of results supports both frameworks in different ways.

We found that participants learned fewer words overall in conditions that included 1x4 trials that encouraged hypothesis-testing. The learning decrement stemmed primarily from poorer performance on "unprobed" items that had been presented only on 4x4 trials. This pattern suggests that any explicit processes employed on the 1x4 hypothesis-testing trials did indeed interfere with learning about the set of items that participants were not explicitly asked about. However, there are two important qualifications to this conclusion. The first is that the effect was transient – as noted above, a second experiment with more training trials found close to ceiling performance regardless of whether the training included 1x4 trials or only 4x4 trials. Second, some items clearly benefited from hypothesis-testing. When participants responded correctly to the 1x4 probe trial, they were far more likely to respond correctly again to that same item in the 18-AFC test than when they responded incorrectly or if the item did not appear on a 1x4 trial. This boost in performance is congruent with the idea that "testing" or practice with information retrieval can produce greater learning benefits than continued studying (Roediger & Karpicke, 2006).

One recent finding in the retrieval may help reconcile the tension between these two findings. Finn & Roediger (2013) found that practicing retrieval of

previously learning associations actually impaired subsequent learning of new associations. This finding may be particularly relevant to the cross-situational word learning paradigm that we employed because the 1x4 hypothesis-testing trials were distributed throughout the training. Thus, each 1x4 trial may have presented both an opportunity for retrieval-enhanced consolidation, benefitting learning of correctly-retrieved items and for retrieval-induced information neglect, interfering with learning of non-retrieved items.

Thus, hypothesis-testing may both facilitate and interfere with cross-situational word learning. However, there are some important caveats to aligning the current paradigm with the literature on effects of retrieval. One is that retrieval practice has been demonstrated to be selectively beneficial for delayed testing, while continued studying produces stronger learning results on immediate testing, which more closely aligns with the test in the current paradigm (Roediger & Karpick, 2006). The second is that much of the literature on effects of retrieval has been conducted using various paired-associates task. These tasks differ from the cross-situational word learning paradigm in important ways. Pairs to be associated are generally presented concurrently and unambiguously and are often in the same modality (two printed words, or a printed word and a picture). The current cross-situational word learning paradigm, however, was specifically designed to model the ambiguity inherent in referential uttterances, where there are typically many referents available in the environment to be associated with any particular word or statement. The benefits of retrieval do generalize beyond paired-associates tasks, as demonstrated by Roediger & Karpicke (2006), who employed more educationally-relevant materials (prose passages to study and a free-recall test). However, more research specifically using the cross-situational word learning paradigm will be required to understand retrieval and learning dynamics in that context.

4.2 Hypotheses in the context of statistics

Other findings from the current study demonstrate that hypothesis during word learning are formed and tested within the context of aggregating statistics, not as an isolated process. Importantly for theories of word learning, we did not find that an incorrect hypothesis about a label's referent meant that it was impossible for participants to link the label to the correct referent by the end of training (Medina et al., 2011). Rather, participants learned such labels at a rate equal to labels that were not probed explicitly during training.

Additionally, we found that the accuracy of participants' hypotheses during training could be predicted by the co-occurrence statistics for the probed label with the target and foil objects presented on the trial. These two findings suggest that explicit hypotheses are likely the *result* of statistical aggregation, rather than being the only means by which learners collect information about potential word meanings. When participants were presented with 1x4 trials but did not have to make an overt response the learning patterns looked similar to the overt response

condition but the effects were smaller. This suggests that the structure of the 1x4 trials may encourage covert hypothesis-testing (or other similar processes) but that the effects of such processing are less strong than when an actual response is made. As discussed above, once explicit hypotheses are formed and tested they are likely to influence the learning process in complex ways - facilitating consolidation of correct information and potentially interfering with the encoding or consolidation of new information.

The current studies provide a novel paradigm for testing interactions between two well-supported mechanisms of word learning – statistical aggregation across instances and hypothesis-testing within an instance. Additional research using such an interactive perspective will broaden our understanding of the dynamics of learning in both adult and developmental populations.

References

Arciuli, Joanne, & Simpson, Ian C. (2012). Statistical learning is lasting and consistent over time. *Neuroscience letters*, *517*(2), 133-135.

Finn, Brigid, & Roediger III, Henry L. (2013). Interfering effects of retrieval in learning new information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(6), 1665.

Horst, Jessica S. & Samuelson, Larissa K. (2008). Fast mapping but poor retention in 24-month-old infants. *Infancy, 13*, 128-157. ISSN 1532-7078

Roediger III, Henry L. & Karpicke, Jeffrey D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17(3)*, 249-255.

Koehne, Judith, Trueswell, John C., & Gleitman, Lila 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.

Markman, Ellen M. (1990). Constraints children place on word meanings. *Cognitive Science*, *14*, 57-77.

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

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning taking memory tests improves long-term retention. *Psychological science*, *17*(3), 249-255.

Romberg, Alexa R. & Yu, Chen (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.

Romberg, Alexa R. & Yu, Chen (2014). Interactions between statistical aggregation and hypothesis testing during word learning. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 1311-1316). Austin, TX: Cognitive Science Society

Saffran, Jenny R. (2002). Constraints on statistical language learning. *Journal of Memory and Language*, *47*(1), 172-196.

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

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

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

Trueswell, John C., Medina, Tamara N., Hafri, Alon, & Gleitman, Lila R. (2013). Propose but verify: Fast mapping meets cross-situational word learning. *Cognitive Psychology*, *66*, 126-156.

Vlach, Haley A., & Johnson, Scott P. (2013). Memory constraints on infants' cross-situational statistical learning. *Cognition*, *127*, 375-382. doi: 10.1016/j.cognition.2013.02.015

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

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

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

Yurovsky, Daniel, Fricker, Damian, Yu, Chen & Smith, Linda B.(2014). The role of partial knowledge in statistical word learning. *Psychonomic Bulletin & Review, 21,* 1-22.

Yurovsky, Daniel, Yu, Chen, & Smith, Linda B. (2012). Statistical speech segmentation and word learning in parallel: Scaffolding from child-directed Sspeech. *Frontiers in Psychology*, *3*, 374. doi:10.3389/fpsyg.2012.00374