An Integrative Account of Cross-Situational Learning

Daniel Yurovsky and Michael C. Frank

Department of Psychology, Stanford University

Author Note

Please address correspondence to:

Daniel Yurovsky

Jordan Hall (Building 420)

Stanford University

450 Serra Mall

Stanford, CA 94305

Email: yurovsky@stanford.edu

Word Count: 4879

References: 47

Abstract

Co-occurrence statistics between words and objects across situations are a powerful information source for language learners. But, there is considerable debate about how we learn from them. While some theories hold that we accumulate graded, statistical evidence about multiple referents for each word, others suggest that we track only a single candidate referent. In two large-scale experiments with adults, we show that neither account is sufficient: cross-situational learning involves elements of both. Further, the empirical data are captured by a computational model that formalizes how memory and attention interact with both single-referent tracking and statistical accumulation. Together, these data unify opposing positions in a complex debate and underscore the value of understanding the interaction between computational and algorithmic levels of explanation.

Keywords: statistical learning, word learning, language acquisition, cognitive development

An Integrative Account of Cross-Situational Learning

Natural languages are richly structured. From sounds to phonemes to words to referents in the world, statistical regularities characterize the units and their connections at every level. Adults, children, and even infants are sensitive to these statistics, leading to a view of language acquisition as a parallel, possibly implicit, process of statistical extraction (Saffran, Aslin, & Newport, 1996; Gómez & Gerken, 2000). Recent experiments across a number of domains, however, show that human statistical learning may be significantly more limited than previously believed (Johnson & Tyler, 2010; Yurovsky, Yu, & Smith, 2012; Trueswell, Medina, Hafri, & Gleitman, 2013).

We focus here on the use of statistical regularities to learning the meanings of concrete nouns (known as cross-situational word learning; Pinker, 1989; Siskind, 1996; Yu & Smith, 2007). Because words' meanings are reflected in the statistics of their use across contexts, learners could discover the meaning of the word "ball" (for instance) by noticing that while it is heard across many ambiguous contexts, it often accompanies play with small, round toys. A growing body of experiments shows that adults, children, and infants are sensitive to such co-occurrence information, and can use it to map words to their referents (Yu & Smith, 2007; Smith & Yu, 2008; Scott & Fischer, 2012; Vlach & Johnson, 2013).

At Marr's (1982) computational level, information about a word's meaning can thus be extracted from the environmental statistics of its use (Frank, Goodman, & Tenenbaum, 2009). However, this computation can be approximated by a number of different algorithmic-level mechanisms, and the one available to humans is under significant debate. Do human learners really maintain a representation of word-object co-occurrences? Some evidence suggests that humans are indeed gradual, parallel accumulators of statistical regularities about the entire system of word-object co-occurrences, simultaneously acquiring information about multiple candidate referents for the same word (Vouloumanos, 2008; McMurray, Horst, & Samuelson, 2012; Yurovsky, Fricker, Yu, & Smith, 2014). Other

evidence suggests that statistical learning is a focused, discrete process in which learners maintain a single hypothesis about the referent of any given word. This referent is either verified by future consistent co-occurrences or instead rejected, "resetting" the learning process (Medina, Snedeker, Trueswell, & Gleitman, 2011; Trueswell et al., 2013). While both of these algorithmic-level solutions will, in the limit, produce successful word-referent mapping, they will do so very different rates. In particular, if learners track only a single candidate referent for each word, it may be necessary to posit additional biases and constraints on learners in order for human-scale lexicons to be learned in human-scale time from the input available to children (Vogt, 2012; Reisenauer, Smith, & Blythe, 2013).

To distinguish between these two accounts, previous experiments exposed learners to words and objects in which co-occurrence frequencies indicated several high-probability referents for the same word. At the group level, participants in these experiments showed gradual learning of multiple referents for the same word (e.g. Vouloumanos, 2008; Yurovsky, Yu, & Smith, 2013); but gradual, parallel learning curves can be observed at the group level even if individuals are discrete, single-referent learners (Gallistel, Fairhurst, & Balsam, 2004; Medina et al., 2011). Experiments measuring the same learner at multiple points—a stronger test—have produced mixed results. In some cases, learners showed clear evidence of tracking multiple referents for each word, suggesting a distributional approximation mechanism at the algorithmic level (Smith, Smith, & Blythe, 2011; Yurovsky, Smith, & Yu, 2013; Dautriche & Chemla, in press). In other experiments, however, learners appear to track only a single candidate referent, and to restart from scratch if their best guess is wrong (Medina et al., 2011; Trueswell et al., 2013).

These mixed results expose a fundamental gap in our understanding of the mechanisms humans use to encode and track environmental statistics critical for learning language. Evidence for each account is separately compelling, but neither account can explain the outcomes of experiments used to support the other. Because previous experiments differ along a number of dimensions—e.g., methodology, stimuli, timing, and

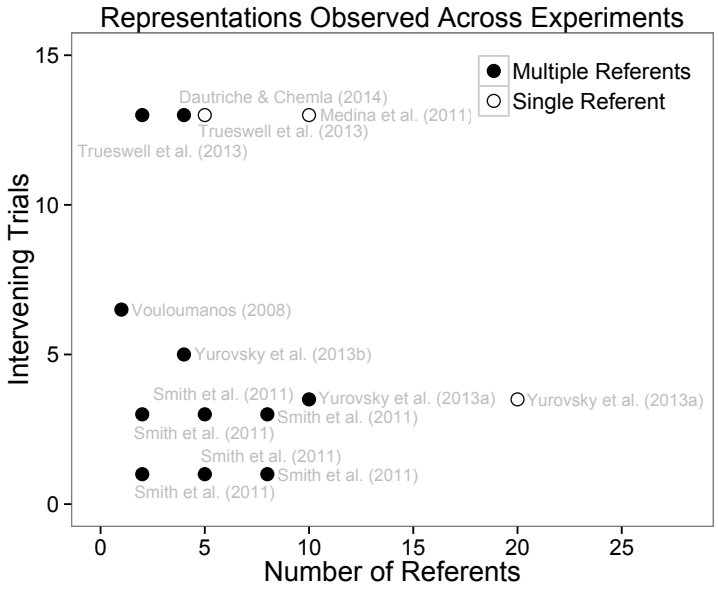

Figure 1. Results of previous experiments investigating representations for cross-situational learning. These experiments vary along a number of dimensions, but two primary dimensions appear to predict whether multiple-referent tracking is observed: Number of referents on a trial, and the interval between trials for the referent.

precision of measurement—it has been difficult to integrate them to understand why cross-situational learning sometimes appear distributional and sometimes appear discrete (see Yurovsky et al., 2014, for a review). We propose that differences in task difficulty may explain diverging results across experiments. Two salient dimensions vary across previous studies: ambiguity of individual learning instances, and the interval between successive exposures to the same label (Figure 1). As attentional and memory demands increase, learners may shift from statistical accumulation to single-referent tracking (Smith et al., 2011; Trueswell et al., 2013). We present a strong test of this hypothesis, adapting a paradigm first introduced by Bower and Trabasso (1963) to study the information learners store in concept identification. We parametrically manipulated both the ambiguity of individual learning trials and the interval between them and measured multiple-referent tracking at the individual-participant level.

Even at the maximum difficulty tested, learners tracked multiple referents for each word rather than a single referent, providing strong evidence against a qualitative shift from statistical accumulation to single-referent tracking. However, the data also show that learners encode the referents with differing strength, remembering much better the one that they hypothesized to be the correct referent. Thus, each previous account appears to be partially correct. To clarify how these two accounts are related, we implement both the single-referent and statistical accumulation accounts as formal models, as well as an integrative model that subsumes both as special cases along a continuum. Only the integrative model can account for the data. We then show that this model makes nearly perfect parameter-free predictions for a follow-up experiment designed to determine the extent to which learners encode mappings rather than individual words and objects. We conclude that cross-situational word learning is best characterized by an integrative account: Learners track both a single target referent and an approximation to the co-occurrence statistics, but the strength of this approximation varies with the complexity of the learning environment (Figure 2).

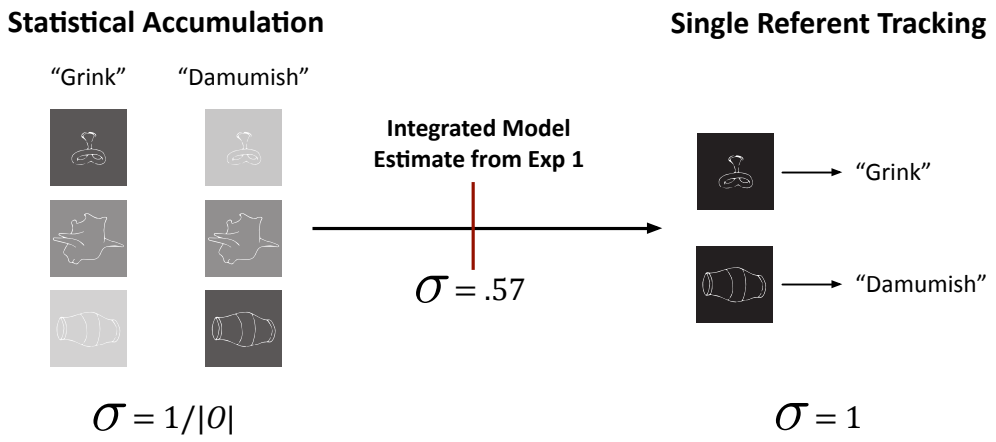

Figure 2. A representation of the continuum between the statistical accumulation and single-referent tracking models as learners' attention is varied from evenly distributed ($\sigma = \frac{1}{|O|}$) to focused on a single referent ($\sigma = 1$), as well as the best-fitting integrated model's position along this continuum.

Experiment 1

We designed Experiment 1 to estimate learners' memory for both their single best hypothesis about the correct referent of a novel word and their additional statistical knowledge as demands on attention and memory varied. Participants saw a series of individually ambiguous word learning trials in which they heard one novel word, viewed multiple novel objects, and made guesses about which object went with each word. To succeed, participants needed to encode at least one of the objects that co-occurred with a word, remember it until their next encounter with that word, and check whether that same object was again present. If participants encoded exactly one object, they would succeed only when their initial hypothesis was correct. However, the more *additional* objects participants encoded on their first encounter with a word, the greater their likelihood of succeeding even if their initial hypothesis was incorrect.

Rather than allowing chance to determine whether participants held the correct hypothesis on their first exposure to a novel word, the set of novel objects presented on the second exposure to each word was constructed based on participants' choices. On *Same* trials, the participant's hypothesized referent was pitted against a set of novel competitors. In contrast, on *Switch* trials, one of the objects the participant had previously *not* hypothesized was pitted against a set of novel competitors (see Figure 3). Logically, either a single-referent tracking or a statistical accumulation mechanism will succeed on Same trials. However, only statistical accumulation of information about non-target items can allow participants to succeed on Switch trials.

Method

Participants. Experiment 1 was posted to Amazon Mechanical Turk as a set of Human Intelligence Tasks (HITs) to be completed only by participants with US IP addresses that paid 30 cents each (for a detailed comparison of laboratory and Mechanical Turk studies see Crump, McDonnell, & Gureckis, 2013). Ninety HITs were posted for each

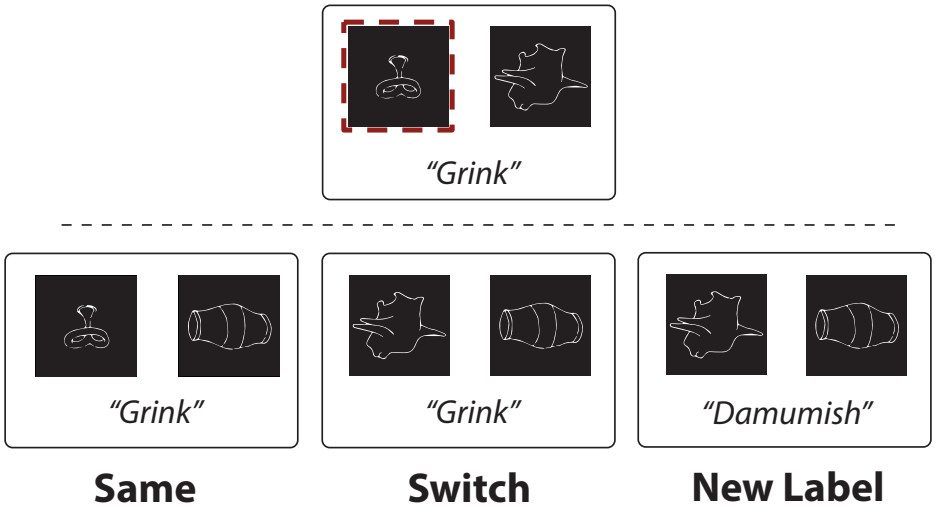

Figure 3. A schematic of the experimental trials seen by participants in Experiments 1 and 2. On their first exposure to each novel word, participants were asked to guess its correct referent. In Experiment 1, the second trial for each word was either a Same trial—the set of referents contained the participant's previous hypothesis, or a Switch trial—the set of referents contained one the participant had previously *not* hypothesized. In Experiment 2, Switch trials were replaced with New Label trials that showed same set of referents but a played a novel word. The number of referents on the screen and the interval between successive exposures to the same word varied across conditions.

of the 16 Referent x Interval conditions for a total of 1440 paid HITs. If a participant completed the experiment more than once, he or she was paid each time but only data from the first HIT completion was included in the final data set (excluded 180 HITs). In addition, data was excluded from the final sample if participants did not give correct answers for familiar trials (64 HITs, see Design and Procedure). The final sample thus comprised 1,196 unique participants, approximately 75 participants per condition (range: 71-81).

Stimuli. Stimuli for the experiment consisted of black and white pictures of familiar and novel objects and audio recordings of familiar and novel words. Pictures of 32 familiar objects spanning a range of categories (e.g. squirrel, truck, tomato, sweater) were

drawn from the set constructed by Snodgrass and Vanderwart (1980). Pictures of distinct but difficult to name novel objects were drawn from the set of 140 first used in Kanwisher, Woods, Iacoboni, and Mazziotta (1997). For ease of viewing on participants' monitors, pixel values for all pictures were inverted so that they appeared as white outlines on black backgrounds (see Figure 3a). 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 pseudowords obeying the rules of English phonotactics produced using the same speech synthesizer.

Design and Procedure

Participants were exposed to a series of trials in which they heard a word, saw a number of objects, and were asked to indicate their guess as to which object was the referent of the word. After a written explanation of this procedure, participants were given four practice trials to introduce them to the task. On each of these trials, they heard a Familiar word and saw a line drawing of that object among a set of other familiar objects. On the first two trials, participants were asked to find the squirrel, and the correct answer was in the same position on each trial. On the next two trials, participants were asked to find the sweater, and the correct answer switched positions from the first to the second trial (in order to ensure that participants understood the on-screen position was not an informative cue to the correct target). These trials also served to screen for participants who did not have their audio enabled or who were not attending to the task.

After these Familiar 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 each of the eight novel words twice, but the order in which these words were presented and the number of objects seen on the screen were varied across sixteen between-subjects conditions. Participants saw either 2, 3, 4, or 8 Referents on each trial, and the two trials for each word occurred either back-to-back, or

were interleaved between trials for other words for an Interval of 1, 2, 3, or 8. Four of these follow-up trials were Same trials in which the referent that participants selected on the first encounter with that object appeared again amongst the set of objects. The other four were Switch trials in which one of the referents in the set was selected randomly from the objects a participant *did not* select on the previous exposure to that word. All other referents were completely novel on each trial. The number of referents on Familiar trials for each participant matched the number of referents they would see on Same and Switch trials.

Because participants performed this task over the internet, it was important to indicate to them that their click had been registered. Thus, a red dashed box appeared around the object they selected on for 1 second after their click was received. This box appeared around the selected object whether or not it was the "correct" referent.[1]

Results

Do statistical learners encode multiple referents for each word, or do they instead encode only a single hypothesized referent? We compared the distribution of correct responses made by each participant to the distribution expected if participants were selecting randomly (defined by a Binomial distribution with four trials and a probability of success of $\frac{1}{\#Referents}$). The top row of Figure 4 shows participants' accuracies in identifying the referent of each word in all conditions for both kinds of trials (Same and Switch). At all Referent and Interval levels, both for Same and for Switch trials, participants' responses differed from those expected by chance (smallest $\chi^2(4) = 15.07$, all $p < .01$). Thus, learners encode more than a single hypothesis in ambiguous word learning situations, even under high levels of memory and attentional load.

Next, to quantify the effect of each factor on word learning, we fit a mixed-effect logistic regression model to the data from the full dataset (Baayen, Davidson, & Bates,

[1]It is possible that forcing participants to select an object on each trial could have changed their performance. However, control conditions from three previous experiments suggest that empirically this is not the case (Medina et al., 2011; Smith et al., 2011; Trueswell et al., 2013).

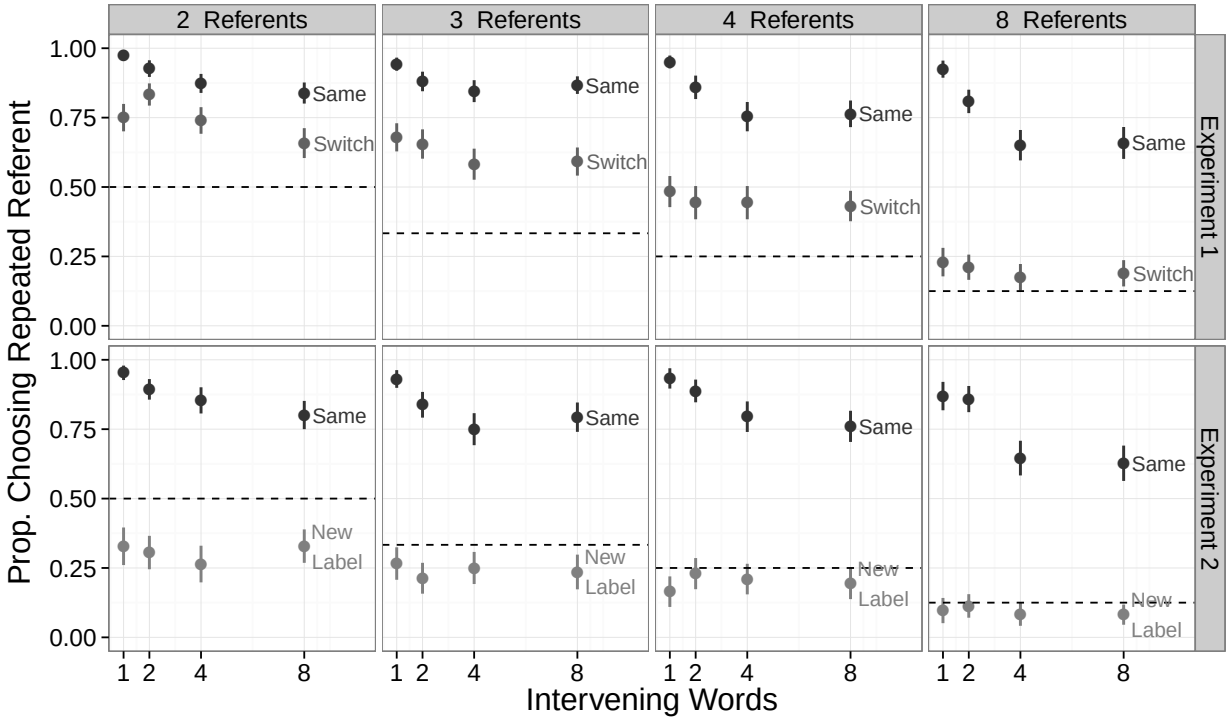

Figure 4. Proportion of repeated referents selected by participants at each combination of number of Referents and Interval on Same and Switch trials in Experiment 1, and Same and New Label trials in Experiment 2. Each datapoint represents 75 participants in Experiment 1 and 50 participants in Experiment 2. Error bars indicate 95% confidence intervals computed by non-parametric bootstrap. Learning in all conditions of Experiment 1 differed from chance and declined mostly due to Interval for Same trials but mostly due to Referents for Switch trials. Experiment 2 Same trials replicated performance in Experiment 1 Same trials, but New Label trials were different from Switch trials in all Referent and Interval conditions.

2008). This analysis showed significant main effects of Number of Referents, Interval, and Trial Type. In addition, the model showed a significant two-way interaction between Referents and Trial Type and a significant three-way interaction between all three factors (Table 1). Thus, while word learning was best at low levels of referential ambiguity and at low memory demands, the decreases in word learning observed on Same and Switch trials

were due to different factors. For Same trials, the number of Referents played a relatively small role in the difficulty of learning, while the Interval between learning and test played a large role. However, for Switch trials, there was relatively little decline in word mapping as Interval increased but a large decline due to number of Referents.

These data suggest that neither the single-referent tracking nor the statistical accumulation account of cross-situational word learning is correct. Although learners did encode multiple referents, they did not encode them all with equal strength. Memory for the hypothesized referent was stronger than for non-hypothesized referents at all referent-set sizes and at all intervals. Further, the difference between them grew with number of referents. Thus, it appears that a new account is necessary that integrates elements of both single-referent tracking and accumulative statistical tracking.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	4.68	0.41	11.45	0.00
log.numPic	-0.55	0.18	-3.00	0.00
log.interval	-0.41	0.19	-2.19	0.03
trialTypeSwitch	-1.44	0.43	-3.34	0.00
log.numPic:log.interval	-0.13	0.09	-1.45	0.15
log.numPic:trialTypeSwitch	-1.04	0.20	-5.32	0.00
log.interval:trialTypeSwitch	0.13	0.20	0.65	0.51
log.numPic:log.interval:trialTypeSwitch	0.20	0.10	2.13	0.03

Before presenting a formal integrative account in the Model section below, we first rule out one other possibility. Because the set of foils for each target referent was distinct, it is possible that participants succeeded on Switch trials by simply selecting the most familiar object regardless of which word they were hearing. If so, these data would be consistent with a slightly amended version of the single-referent tracking account in which learners also have some residual memory for previously-seen objects but have not learned them as word-object mappings. Experiment 2 presents a new learning condition to test this possibility.

Experiment 2

Participants' above-chance accuracy on Switch trials in Experiment 1 provides evidence of their memory for multiple objects, but not necessarily for the formation of referential mappings between the objects and the novel words. To rule out this second possibility, Experiment 2 replaced Switch Trials with New Label trials in which participants saw an object they had previously *not* selected among a set of novel competitors but heard a *New Label*. If the underlying mechanism for tracking multiple referents is pure referent familiarity, New Label trials should produce similar responses to the Switch trials in Experiment 1. In contrast, if success on Switch trials was due to a mapping between words and referents, New Label trials should show a different pattern of performance.

Method

Participants. As in Experiment 1, participants for Experiment 2 were recruited from Amazon Mechanical Turk under the constraint that they had a US IP address. Each HIT paid 30 cents for completion. Sixty HITs were posted for each of the sixteen Referent x Interval conditions for a total of 960 paid HITs. Participants were again paid for multiple HITs, but only data from their first was included in the final set (excluded 100 HITs). In addition, data was again excluded from the final sample if participants did not give correct answers for familiar trials (60 HITs). The final sample thus comprised 803 unique

participants, approximately 50 participants per condition (range: 41–55).

Stimuli, Design, and Procedure

All aspects of the Stimuli, Design, and Procedure of Experiment 2 were identical to those of Experiment 1 except for the construction of New Label trials. On these trials, the set of candidate referents was the same as on Switch trials in Experiment 1, but the word was novel (Figure 3).

Results

Participants showed robust evidence of learning mappings (rather than simply tracking familiar objects). Whereas participants on Same trials were more likely than predicted by chance to select a referent they had previously seen but not guessed, participants in New Label trials were, in many cases, *less* likely than predicted by chance to select these same referents. Further, in all Referent and Interval conditions, performance on New Label trials differed significantly from performance on comparable Switch trials. That is, these participants recognized these referents from their first exposure, and further recognized that they did not co-occur on their previous exposure with the label they heard at test (bottom row of Figure 4).

In addition, a mixed-effects logistic regression largely reproduced the patterns observed in Experiment 1—word learning accuracies on Same trials declined predominantly due to Interval between learning and test, and very little due to the number of Referents. New Label trials were driven almost entirely by the number of Referents—as was the case with Switch trials in Experiment 1.

Taken together, these data are strong evidence that neither the single-referent tracking nor the statistical accumulation account of cross-situational word learning can be correct. Instead, it appears that cross-situational word learning is best characterized by a combination of both of these mechanisms. In the next section, we formalize this idea.

Predictor	Estimate	Std. Error	z value	p value	
Intercept	3.97	0.25	16.01	<0.0001	***
Log(Referents)	-.48	.09	-4.85	<0.001	***
Log(Interval)	-.60	.07	-8.57	<0.001	***
New Label Trial	-4.03	.28	-14.09	<0.0001	***
Log(Referents)*Switch	-.24	.12	-1.98	<.05	*
Log(Interval)*New Label	.58	.08	7.13	<0.0001	***

Table 1

Predictor estimates with standard errors and significance information for a logistic mixed-effects model predicting word learning in Experiment 2. The model was specified as `Correct ~ Log(Referents) * TrialType + Log(Interval) * TrialType + (TrialType | subject)`

Model

We begin by describing the computational-level learning problem posed by Experiment 1 using the model developed by Frank et al. (2009). In this framework, the learner observes a set of situations S with the goal of determining the lexicon of word-object mappings L that produced them $P(L|S)$. Each of these situations consists of two observed variables: objects (O) and words (W). In addition, situations implicitly contain an additional hidden variable: an intention (I) by the speaker to refer to one of the objects. Thus, from the set of objects present in the situation, the speaker first chooses one to refer to and then produces its word label.

We can use Bayes' rule to determine the inferential computation the learner must perform:

$$P(L|S) \propto P(S|L)\, P(L) \tag{1}$$

Each trial in our experiment consisted of a word and a set of objects. We model the unobserved link between the word and one of the objects as I, the referential intention. Thus, the probability of a lexicon is given as the joint probability of observing all of the words, objects, and intentions given that lexicon times the lexicon's prior probability:

$$P(L|S) \propto \prod_{s \in S} P(W_s, I_s, O_s, |L)\ P(L) \tag{2}$$

Because the referential intention mediates the relationship between words and objects (Frank et al., 2009), we can rewrite this using the chain rule as:

$$P(L|S) \propto \prod_{s \in S} P(W_s|I_s, L)\ P(I_s|O_s)\ P(L) \tag{3}$$

To make predictions from this model, we need to define the probabilities in Equation 3. Following Frank et al. (2009), we propose that the word (W) used to label the intended referent on each trial is chosen uniformly from the set of all words in the lexicon for that object (L_o). In addition, we propose a simple parsimony prior for the lexicon. A priori, the larger the set of words in the lexicon that refer to the same object O, the lower the probability of that lexicon: $P(L_o) \propto \frac{1}{|L_o|}$.

We can then take this computational-level description of the problem and add cognitive constraints to understand how the patterns observed in our data arise from the interaction of learning mechanisms, attention, and memory (see e.g. Frank, Goldwater, Griffiths, & Tenenbaum, 2010; Shi, Griffiths, Feldman, & Sanborn, 2010). We start by describing how participants allocate their attention on each learning trial, a critical point of difference between the two different accounts of cross-situational learning.

In this modeling framework, the most convenient place to integrate attention is in defining he learner's beliefs about $P(I|O)$, the probability of the speaker choosing to refer to each object in the set. One possibility is to let each object be equally likely to be the intended referent, implementing parallel Statistical Accumulation as in Frank et al. (2009).

Alternatively, the learner could place all of the probability mass on one hypothesized referent – implementing a Single Referent tracking strategy. A more flexible alternative is to assign some probability mass σ to the hypothesized referent, and divide the remainder evenly among the remaining objects: $\frac{1-\sigma}{|O|-1}$. This Integrated model subsumes the other two as special cases: At $\sigma = 1$, it becomes a Single Referent tracker, and at $\sigma = \frac{1}{|O|}$, it becomes a parallel Statistical Accumulator (Figure 3b). We note that there is some debate about the mechanisms that give rise to attentional limitations – for instance whether attention is allocated in discrete slots or as a continuous resource (Luck & Vogel, 1997; Vul, Hanus, & Kanwisher, 2009; Wei, Wang, & Wang, 2012). In our formulation, attention is treated more like a continuous resource, but this a matter of convenience rather than an important theoretical commitment. For our purposes, the important question is to what extent attention is focused on the single target referent, and a continuous implementation allows parameter-estimation to answer this question.

Next, we model how learners' memories for observed situations decay over time. We follow previous memory researchers by formalizing memory for a lexical entry as a *power function* of the interval between successive exposures, where γ scales the strength of initial encoding and λ defines the rate of decay (Anderson & Schooler, 1991). As with attention-allocation, we note that there a number of successful models of the underlying mechanisms that give rise to phenomena like the power-law observed in human memory (e.g. Murdock, 1982; Shiffrin & Steyvers, 1997). However, as with attention, the critical aspect for modeling this data is to be consistent with the broader dynamics of human memory, rather than with determining which model can best account for these dynamics.

$$M(L_o) = \gamma\, L_o\, t^{-\lambda} \tag{4}$$

Finally, we provide a choice rule for describing how learners select at test among the objects on each trial. We propose that learners choose the correct referent with probability proportional to their memory for its lexical entry, and otherwise choose randomly among

the set of referents.[2] We use this rule because all of the foils on both Same and Switch trials are novel, and thus should have no trace in memory.

$$P(Correct) = M(L_o) + \frac{1 - M(L_o)}{|O|} \tag{5}$$

All 3 models—Statistical Accumulation, Single Referent, and Integrated—were fit to the data from Experiment 1 at the individual-participant level. While the Single Referent and Statistical Accumulation models capture some of the structure in the data, each leaves significant variance unexplained. The Single Referent Model cannot predict above-chance performance on Switch trials, and the Statistical Accumulation model cannot predict a difference between the Same and Switch trials. The Integrated model, however, predicts 95% of the variance in the data, and significantly outperforms the other models in BIC comparisons as well—a metric that trades off its superior performance against its one additional parameter. Table 3 shows statistics of fit for all models.

Model	Log Likelihood	BIC	E1 r^2	E1+2 r^2
Statistical Accumulation	-5924	11863	.32	.67
Single Referent	-5337	10690	.83	.77
Integrated	**-4854**	**9733**	**.95**	**.98**

Table 2

Likelihood and Correlation measures for models on Experiments 1 and 2. The Integrated model outperformed both of the individual accounts on all measures.

In addition, we can use the models presented above, with parameters estimated from Experiment 1, to make parameter-free predictions about the data observed in Experiment 2. As before, the Single Hypothesis and Parallel Accumulation models predict some of the

[2]This formulation is equivalent to using Luce's (1959) Choice Axiom in which the target has strength $M(L_o) + \frac{1-M(L_o)}{|O|}$ and each foil has strength $\frac{1-M(L_o)}{|O|}$.

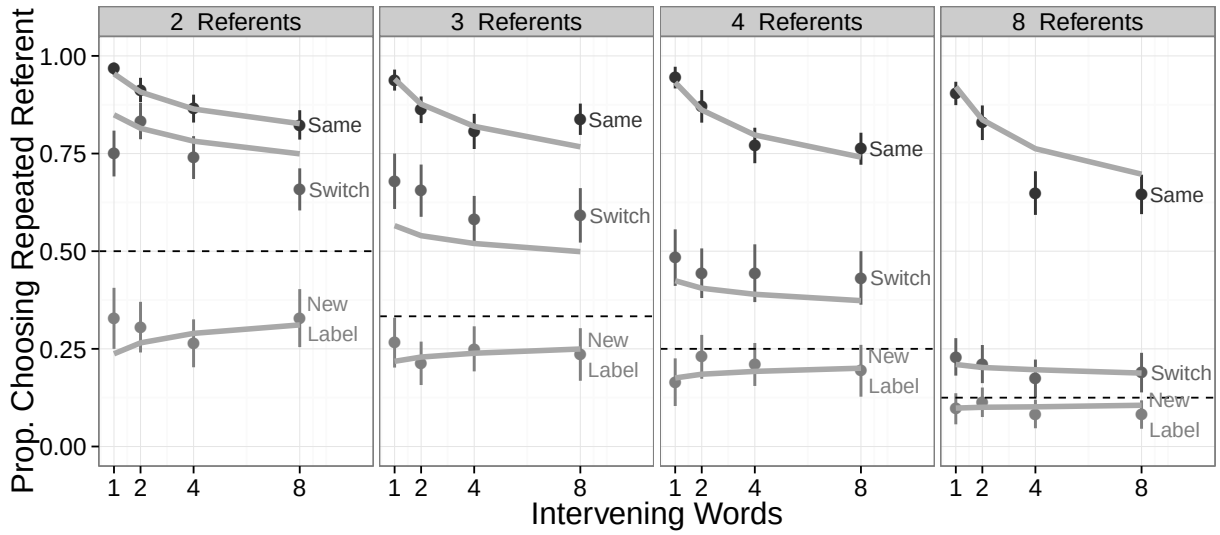

Figure 5. Predictions of the Integrated model for all conditions in Experiment 1 and 2. The model was able to account for 98% of the variance in the data, significantly outperforming both the Single Hypothesis and Parallel Accumulation models.

variance in the new data, but leave much unexplained. With parameter values fixed to those estimated for Experiment 1, the Integrated model makes near-perfect predictions about the new data—including the New Label condition—explaining 98% of the combined variance in the data from Experiments 1 and 2 (Table 3). Figure 5 presents model predictions for all of the data collected. Taken together, Experiments 1 and 2 and the computational model thus provide strong evidence that adults track not only a single hypothesis for the most likely referent of a novel word, but also some approximation to distributional statistics that becomes less precise as the number of referents increases.

General Discussion

For an ideal learner, word-object co-occurrence statistics contain a wealth of information about meaning. But how is this information used by human learners? One possibility is that learning is fundamentally statistical, and we gradually accumulate distributional information across situations. Another possibility, however, is that we track

only a single, discrete hypothesis at any time. While each of these accounts has some support in prior work, neither is consistent with all of the extant data.

Our results here suggest an explanation: The degree to which learners represent statistical information depends on the complexity of the learning situation. When there are many possibilities, learners represent little about any other than the one that is currently favored; when there are few, learners represent more. This account does not depend on positing multiple, discrete learning systems. Instead, the tradeoff between the most likely hypothesis and the alternatives emerges from graded constraints on memory and attention. Consistent with this account, when we manipulated the cognitive demands of a cross-situational word learning paradigm, we found a gradual shift in the fidelity with which alternatives were represented.

The graded shift in representation we found was well-described by an ideal learning model, but only when this model was modified to take into account psychological constraints on attention and memory (Kachergis, Yu, & Shiffrin, 2012; Smith & Yu, 2013; Vlach & Johnson, 2013; Yurovsky et al., 2014). This modeling framework allowed us to estimate the effects of these constraints on learning to find the model that best fit the data—one that is intermediate between the two extreme poles of parallel statistical accumulation and single-referent tracking. This unifying account provides a route by which both hypotheses and sensitivity to statistics can make complementary contributions to word learning (Waxman & Gelman, 2009; Kachergis, Yu, & Shiffrin, 2013).

The shift from a computational to a psychologically-constrainted level of description was critical in capturing the pattern of human performance in our task (Marr, 1982; Frank et al., 2010; Yurovsky et al., 2012). For the current model, we chose one principled instantiation of cognitive limitations based on previous work, but there may be other consistent proposals. Indeed, recent work from Yu and Smith (2012) suggests that human performance observed in cross-situational learnings task can be consistent with a number of seemingly quite different models that can mimic each other (see also Townsend, 1990;

Townsend & Wenger, 2004). They note that modeling choices that some may consider peripheral to the central learning mechanism—e.g., attentional allocation, memory, choice rule—can be varied to produce many different patterns of learning. They thus propose a decomposition of cross-situational learning into a number of sub-processes that are studied and fixed independently. We agree with their central argument about the importance of these kinds of processes to learning, but take a different approach here. First, we fit a large set of parametrically-varying data that imposes strong constraints on model parameters. Second, we prevented overfitting by fixing model parameters using Experiment 1, and making parameter-independent predictions about human performance that were supported in Experiment 2. This approach allowed us to gain insight about both the central learning mechanism and the "peripheral" processes that determine human performance.

Our work shows how ideal learning models can inform psychological accounts of statistical learning more broadly. Although we focused here on noun learning, our results are relevant to work on many aspects of early language, including phonetic category learning, speech segmentation, and grammar learning. In each of these domains, researchers have debated the degree to which learners represent distributional information (Endress, Scholl, & Mehler, 2005; Frank et al., 2010; McMurray, Kovack-Lesh, Goodwin, & McEchron, 2013). We suggest a synthesis. Learning is fundamentally distributional, but the fidelity of learners' distributional estimates depends critically on their limited attention and memory.

Acknowledgements

We are grateful to to Linda Smith, Elika Bergelson, Molly Lewis, Ann Nordemeyer, and all of the members of the Language and Cognition Lab for their feedback on this project. This work was supported by a NIH NRSA F32HD075577 to DY as well as a grant from the Merck Scholars Foundation to MCF.

References

Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological Science*, *2*, 396–408.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412.

Bower, G., & Trabasso, T. (1963). Reversals prior to solution in concept identification. *Journal of Experimental Psychology*, *66*(4), 409-418.

Crump, M. J. C., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PLOS ONE*, *8*(3), e57410.

Dautriche, I., & Chemla, E. (in press). *Journal of Experimental Psychology: Learning, Memory, and Cognition.*

Endress, A. D., Scholl, B. J., & Mehler, J. (2005). The role of salience in the extraction of algebraic rules. *Journal of Experimental Psychology: General*, *134*(3), 406–419.

Frank, M. C., Goldwater, S., Griffiths, T. L., & Tenenbaum, J. B. (2010). Modeling human performance in statistical word segmentation. *Cognition*, *117*, 107–125.

Frank, M. C., Goodman, N., & Tenenbaum, J. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, *20*, 578–585.

Gallistel, C. R., Fairhurst, S., & Balsam, P. (2004). The learning curve: Implications of a quantitative analysis. *Proceedings of the National Academy of Sciences*, *101*, 13124–31.

Gómez, R. L., & Gerken, L. (2000). Infant artificial language learning and language acquisition. *Trends in Cognitive Sciences*, *4*, 178–186.

Johnson, E. K., & Tyler, M. D. (2010). Testing the limits of statistical learning for word segmentation. *Developmental Science*, *13*, 339–345.

Kachergis, G., Yu, C., & Shiffrin, R. M. (2012). An associative model of adaptive inference

for learning word-referent mappings. *Psychonomic Bulletin & Review*, *19*, 317-324.

Kachergis, G., Yu, C., & Shiffrin, R. M. (2013). Actively learning object names across ambiguous situations. *Topics in Cognitive Science*, *5*(1), 200–213.

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.

Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis*. New York, NY: Wiley.

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, *390*(6657), 279–281.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. New York, NY: W. H. Freeman.

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*, 831–877.

McMurray, B., Kovack-Lesh, K. A., Goodwin, D., & McEchron, W. (2013). Infant directed speech and the development of speech perception: Enhancing development or an unintended consequence? *Cognition*, *129*(2), 362–378.

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

Murdock, B. B. (1982). A theory for the storage and retrieval of item and associative information. *Psychological Review*, *89*, 609–626.

Pinker, S. (1989). *Learnability and cognition: The acquisition of argument structure*. Cambridge, MA.: MIT Press.

Reisenauer, R., Smith, K., & Blythe, R. A. (2013). Stochastic dynamics of lexicon learning in an uncertain and nonuniform world. *Physical Review Letters*, *110*(25), 258701.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*, 1926–1928.

Scott, R. M., & Fischer, C. (2012). 2.5-year-olds use cross-situational consistency to learn verbs under referential uncertainty. *Cognition*, *122*, 163–180.

Shi, L., Griffiths, T. L., Feldman, N. H., & Sanborn, A. N. (2010). Exemplar models as a mechanism for performing Bayesian inference. *Psychonomic Bulletin & Review*, *17*, 443–464.

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM - retrieving effectively from memory. *Psychonomic Bulletin & Review*, *4*, 145–166.

Siskind, J. M. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, *61*, 39–91.

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

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

Smith, L. B., & Yu, C. (2013). Visual attention is not enough: Individual differences in statistical word-referent learning in infants. *Language, Learning, and Development*, *9*, 25–49.

Snodgrass, J. G., & Vanderwart, M. (1980). A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Human Learning and Memory*, *6*(2), 174–215.

Townsend, J. T. (1990). Serial vs. parallel processes: Sometimes they look like Tweedledum and Tweedledee but they can (and should) be distinguished. *Psychological Science*, *1*, 46–54.

Townsend, J. T., & Wenger, M. J. (2004). The serial–parallel dilemma: A case study in a linkage of theory and method. *Psychonomic Bulletin & Review*, *11*, 391–418.

Trueswell, J. C., Medina, T. N., Hafri, A., & Gleitman, L. R. (2013). Propose but verify:

Fast mapping meets cross-situational learning. *Cognitive Psychology*, *66*, 126–156.

Vlach, H. A., & Johnson, S. P. (2013). Memory constraints on infants' cross-situational statistical learning. *Cognition*, *127*(3), 375–382.

Vogt, P. (2012). Exploring the robustness of cross-situational learning under Zipfian distributions. *Cognitive Science*, *36*, 726–739.

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

Vul, E., Hanus, D., & Kanwisher, N. (2009). Attention as inference: Selection is probabilistic; responses are all-or-none samples. *Journal of Experimental Psychology: General*, *138*(4), 540-560.

Waxman, S. R., & Gelman, S. A. (2009). Early word-learning entails reference, not merely associations. *Trends in Cognitive Science*, *13*, 258–263.

Wei, Z., Wang, X. J., & Wang, D. H. (2012). From distributed resources to limited slots in multiple-item working memory: A spiking network model with normalization. *Journal of Neuroscience*, *32*(33), 11228–11240.

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*, 21–39.

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

Yurovsky, D., Smith, L. B., & Yu, C. (2013). Statistical word learning at scale: The baby's view is better. *Developmental Science*, *16*, 959–966.

Yurovsky, D., Yu, C., & Smith, L. B. (2012). Statistical speech segmentation and word learning in parallel: Scaffolding from child-directed speech. *Frontiers in Psychology*, *3*, 374.

Yurovsky, D., Yu, C., & Smith, L. B. (2013). Competitive processes in cross-situational

word learning. *Cognitive Science*, *37*, 891-921.