An integrative account of constraints on cross-situational learning

Daniel Yurovsky *, Michael C. Frank

Department of Psychology, Stanford University, United States

ABSTRACT

Word-object co-occurrence statistics are a powerful information source for vocabulary learning, but there is considerable debate about how learners actually use them. While some theories hold that learners accumulate graded, statistical evidence about multiple referents for each word, others suggest that they track only a single candidate referent. In two large-scale experiments, 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 co-occurrence tracking. Together, the data and model unify opposing positions in a complex debate and underscore the value of understanding the interaction between computational and algorithmic levels of explanation.

© 2015 Elsevier B.V. All rights reserved.

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 have been shown to be sensitive to these statistics, leading to a view of language acquisition as a parallel, possibly implicit, process of statistical extraction (Gómez & Gerken, 2000; Saffran, Aslin, & Newport, 1996). Recent experiments across a number of domains, however, show that human statistical learning may be significantly more limited than previously believed (Johnson & Tyler, 2010; Trueswell, Medina, Hafri, & Gleitman, 2013; Yurovsky, Yu, & Smith, 2012).

We focus here on the use of statistical regularities to learn 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 (Smith & Yu, 2008; Suanda, Mugwanya, & Namy, 2014; Vlach & Johnson, 2013; Yu & Smith, 2007).

Information about a word's meaning can thus be extracted from the environmental statistics of its use (Frank, Goodman, & Tenenbaum, 2009; Siskind, 1996). But this analysis is posed at what Marr (1982) called the "computational theory" level: dealing only with the nature of the information available to the learner. At

the "algorithmic" level—the level of psychological instantiation in the mind of the learner—this idealized statistical computation could be realized in many ways, and the computation human learners actually perform is a topic of significant debate (see e.g., Yu & Smith, 2012).

Do human learners really track and 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 (McMurray, Horst, & Samuelson, 2012; Vouloumanos, 2008; 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 at very different rates. In particular, if learners track a only a single 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 (Blythe, Smith, & Smith, 2010; 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,

* Corresponding author.
  E-mail address: yurovsky@stanford.edu (D. Yurovsky).

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 (Dautriche & Chemla, 2014; Smith, Smith, & Blythe, 2011; Yurovsky et al., 2013). 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 evidence used to support the other. Because previous experiments differ along a number of dimensions—e.g., methodology, stimuli, timing, and precision of measurement—it has been difficult to integrate them to understand why cross-situational learning sometimes appear distributional and sometimes appear discrete (for a review, see Yurovsky et al., 2014).

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 (Fig. 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 test of this hypothesis, adapting a paradigm first introduced in 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; this result constitutes strong evidence against a qualitative shift from statistical accumulation to single-referent tracking. The data also show that learners encode the referents with differing strengths, however, remembering their hypothesized referent much better. Thus, each previous account appears to be partially correct.

To clarify how these two accounts are related, we implemented both single-referent tracking and statistical accumulation as computational models. We also extended these accounts into an integrative model that subsumes both as special cases along a continuum. Only the integrative model accounted for our full dataset.
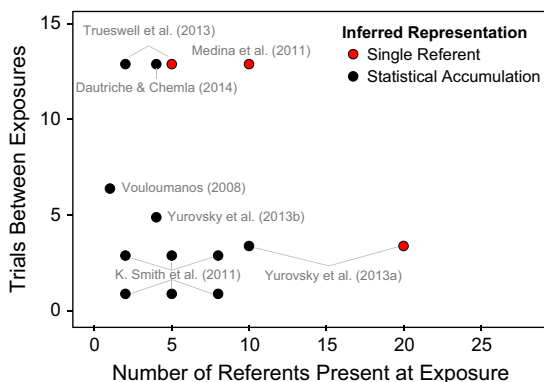
Further, this model was able to make nearly perfect parameter-free predictions for a follow-up experiment that was designed to verify that learners encode mappings rather than individual words and objects. We conclude that cross-situational word learning is best characterized by an integrative account: Learners track both a single target referent and an approximation to the co-occurrence statistics; the strength of this approximation varies with the complexity of the learning environment.

1. 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 Fig. 2). 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 succeed at above-chance levels on Switch trials.

1.1. Method

1.1.1. 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 of the 16 Referent × 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 1196 unique participants, approximately 75 participants per condition (range: 71–81).

1.1.2. 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

Fig. 1. Results of previous experiments investigating representations for cross-situational learning. These experiments vary along a number of dimensions, but two appear to predict whether multiple-referent tracking is observed: the number of referents present on each trial, and the interval between trials for the referent.
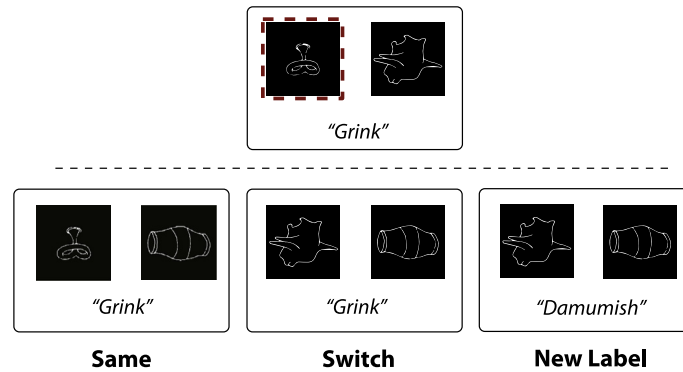
Fig. 2. A schematic of the experimental trials seen by participants in Experiments 1 and 2. On their first exposure to each novel word, participants were asked to guess its correct referent. In Experiment 1, the second trial for each word was either a Same trial—the set of referents contained the participant's 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.

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 Fig. 2). Familiar words consisted of the labels for the familiar objects as produced by AT&T Natural Voices™ (voice: Crystal). Novel words were 1–3 syllable pseudowords obeying the rules of English phonotactics produced using the same speech synthesizer.

1.1.3. 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 that 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 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 s after their click was received. This box appeared

around the selected object whether or not it was the "correct" referent.[1]

1.2. Results

Do statistical learners encode multiple referents for each word, or do they instead encode only a single hypothesized referent? The top row of Fig. 3 shows participants' accuracies in identifying the referent of each word in all conditions for both kinds of trials (Same and Switch). To determine whether participants were learning word-referent mappings, we asked whether these accuracies were significantly different than would be expected by chance. Because these accuracies were estimated from a small number of discrete choices for each participant in each condition, they violate the assumptions of standard continuous analyses like *t*-tests. A better model of chance behavior for this data is a *Binomial* distribution with a probability of success $p = 1/\#Referents$.

To test whether participants' selecting the previously exposed referents more often than predicted by this null model, we fit logistic regressions for each Referent, Interval, and Trial Type combination. These modes were specified as `Correct ~ 1 + offset(logit(1/Referents))`, where the offset encodes the chance probability of success given the number of referents. The intercept term in each of these models captures on a log-odds scale how much more likely participants are to select the correct referent than would be expected by chance. At all Referent and Interval levels, both for Same and for Switch trials, participants chose the correct referent more often than would be expected by chance (smallest $\beta = .393, z = 2.55$, all $ps \leqslant .01$). Thus, learners encoded 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-effects logistic regression model to the data from the full dataset (Baayen, Davidson, & Bates, 2008). All mixed-effects models presented in the paper were implemented in R 3.13 using version 1.1-7 of the `lme4` package. Because of the complexity of the dataset, we constructed models iteratively, with first main effects and then interaction terms added as long as they significantly improved the fit of the model to the data (measured by likelihood comparison tests using χ^2). In addition, as in the comparison to chance above, we used an offset of *logit*(1/*Referents*)

---

[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), and Trueswell et al. (2013).
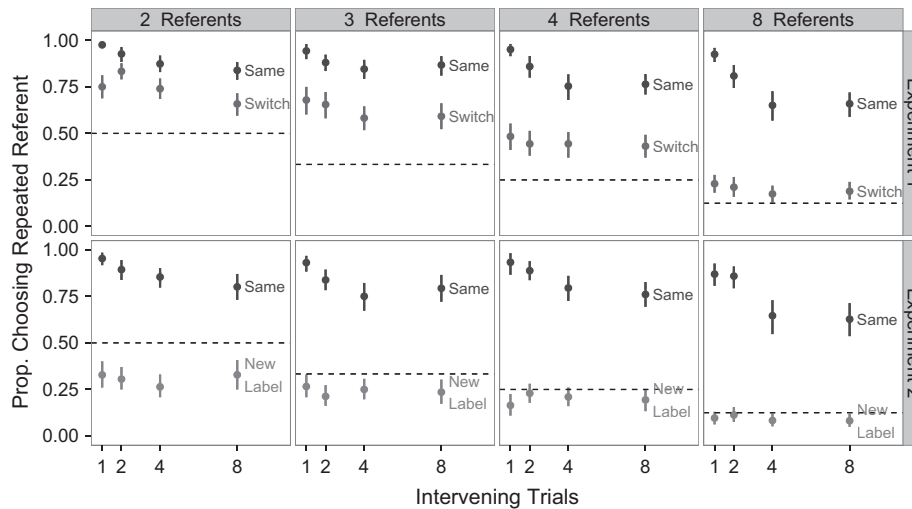
Fig. 3. Proportion of repeated referents selected by participants at each combination of number of Referents and Interval on Same and Switch trials in Experiment 1, and Same and New Label trials in Experiment 2. Each datapoint represents ~75 participants in Experiment 1 and ~50 participants in Experiment 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.

so that each Referents condition was corrected for its different chance performance probability.

This analysis showed a significant Intercept term—indicating globally above-chance performance—as well as main effects of Interval and Trial Type. In addition, the model showed significant two-way interaction between Referents and Trial Type and Interval and Trial Type (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, because performance was comparatively less good at low Intervals, 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.

Table 1
Predictor estimates with standard errors and significance information for a logistic mixed-effects model predicting word learning in Experiment 1. The model was specified as `Correct ~ Log(Referents) * TrialType + Log(Interval) * TrialType + offset(logit(1/Referents)) + (TrialType|subject)`.

Predictor	Estimate	Std. error	z value	p value	
Intercept	4.31	0.28	15.18	<.001	***
Log(Referents)	0.17	0.10	1.66	0.10	.
Log(Interval)	−0.68	0.07	−9.72	<.001	***
Switch trial	−2.16	0.30	−7.22	<.001	***
Log(Referents) * Switch trial	−0.68	0.11	−6.20	<.001	***
Log(Interval) * Switch trial	0.54	0.07	7.34	<.001	***

Before presenting a formal integrative account in the Model section below, we first rule out one other possibility. Because the set of competitors for each target referent was distinct, participants could have succeeded on Switch trials by selecting the most familiar object regardless of which word they were hearing. If so, these data would be consistent with a slightly amended 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.

2. Experiment 2

Participants' above-chance accuracies on Switch trials in Experiment 1 provide 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* (Fig. 2). If success on Switch trials was due purely to referent familiarity, New Label trials should produce similar responses. In contrast, if success on Switch trials was due to a learned mapping between words and referents, New Label trials should show a different pattern of performance.

2.1. Method

2.1.1. 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 × 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).

2.1.2. 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 (Fig. 2).

2.2. Results

Participants showed robust evidence of learning mappings (rather than simply tracking familiar objects). As in Experiment 1, we used logistic regression to determine whether participants chose the previously-seen referent on test trials at above-chance levels. As before, participants selected their perviously guessed referent on Same trials at levels far exceeding chance (smallest $\beta = 1.39, z = 8.17$, all ps <.001). In contrast, on New Label trials—in which a novel label was paired with a previously seen but not guessed referent—participants *never* selected the previous referent at above-chance levels. Further, in the 2 and 3 Referents conditions, they reliably selected the previously seen but not guessed referent at *below* chance levels (largest $\beta = -.32, z = 1.95$, all ps ⩽.05). In the 4 and 8 Referents conditions, participants also selected the previously seen referent less frequently than predicted by chance, but this difference was not statistically reliable.

To determine whether performance on these harder New Label trials was nonetheless different from comparable Switch trials in Experiment 1, we again fit an intercept-adjusted a logistic regression for each Referent x Interval condition, but included a Condition (Switch vs. New Label) term: `Correct ∼ 1 + Condition + offset(logit(1/Referents))`. The Condition term was reliably different from 0 in all conditions, indicating that participants treated Switch and New Label trials differently (smallest $\beta = .74, z = 2.87$, all ps ⩽.001). That is, participants recognized the previous referents on New Label trials from their first exposure to these referents, and further recognized that they had not co-occured on their previous exposure with the New Label (bottom row of Fig. 3). This is strong evidence that participants did indeed encode word-object mappings for non-guessed referents even at the highest number of Referents and at the longest Interval.

In addition, a mixed-effects logistic regression largely reproduced the patterns observed in Experiment 1 (Table 2). Performance on Same trials declined predominantly due to Interval, but not due to number of Referents. In contrast, interval had very little effect New Label trials—as was the case for 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 is correct.[2] Instead, cross-situational word learning is best characterized by a combination of both of these mechanisms. In the next section, we formalize this idea.

3. Model

We begin by describing the computational-level learning problem posed by Experiment 1 using the model developed in 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 $P(L|S)$. We can use Bayes' rule to describe the inferential computation the learner must perform:

Table 2

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) + Log(Interval) * TrialType + offset(logit(1/Referents)) + (TrialType | subject)`.

Predictor	Estimate	Std. error	z value	p value	
Intercept	3.42	0.21	16.10	<.001	***
Log(Referents)	0.32	0.06	5.52	<.001	***
Log(Interval)	−0.60	0.07	−8.22	<.001	***
New Label trial	−4.49	0.19	−23.52	<.001	***
Log(Interval) * New Label trial	0.58	0.08	6.89	<.001	***

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

Each situation 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, speakers first choose an object from the set and then choose a referential label for it. 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 Eq. (2) using the chain rule:

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

To make predictions from this model, we need to define the probabilities in Eq. (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 framework, the most convenient place to integrate attention is in defining the learner's beliefs about $P(I|O)$, the probability of the speaker choosing to refer to each object in the set.[3] 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

---

[2] One alternative explanation remains possible: Perhaps participants track word and referent familiarity independently, and map familiar words to familiar referents and unfamiliar words to unfamiliar referents, without ever linking the two. For an additional control experiment that rules out this explanation, see Appendix A.

[3] A full process model should in principle include two distinct components: a learner's inferred beliefs about a speaker's referential intention and the subsequent decision to allocate attention on the basis of these beliefs. But these processes are indistinguishable in our data, and consequently we collapse them down to a single parameter that controls allocation of attention; future work should distinguish them, however.
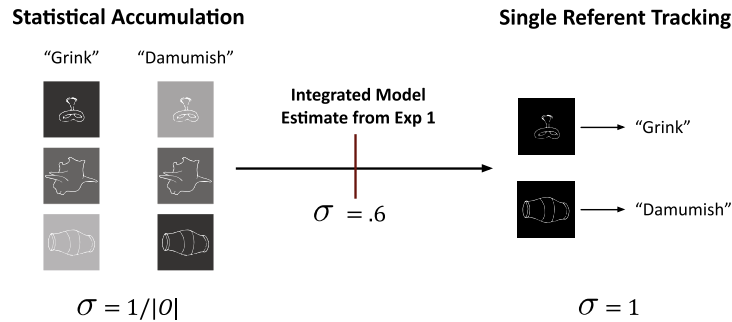
Statistical Accumulation   **Single Referent Tracking**

Fig. 4. A representation of the continuum between the Statistical Accumulation and Single Referent Tracking models as learners' attention is varied from evenly distributed $\left(\sigma = \frac{1}{|O|}\right)$ to focused on a Single Referent ($\sigma = 1$), as well as the best-fitting Integrated model's position along this continuum.

Table 3
Likelihood and Correlation measures for models on Experiments 1 and 2. The Integrated model outperformed both of the individual accounts on all measures.

Model	Log likelihood	BIC	E1 r^2	E1 + 2 r^2
Statistical Accumulation	−6565	13,145	0.33	0.66
Single Referent	−5950	11,915	0.83	0.77
Integrated	−5590	11,203	0.95	0.97

$\sigma = 1$, it is a Single Referent tracker, and at $\sigma = \frac{1}{|O|}$, it is a parallel Statistical Accumulator (Fig. 4).

There is some debate about the mechanisms that give rise to attentional limitations (e.g. Wei, Wang, & Wang, 2012). In our formulation, attention is treated as a continuous resource, but this choice is a matter of convenience rather than a 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 (Anderson & Schooler, 1991). As with attention-allocation, 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). Again, 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. Accordingly, memory for lexical entry L_o decays according to a power function of time t in which γ scales the strength of initial encoding and λ defines the rate of decay.

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

Finally, we provide a choice rule describing how learners select among the objects on each test 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 (Eq. (5)).[4] We use this rule because all of the competitors on both Same and Switch trials were novel, and thus should have no trace in memory.

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

We implemented our models in R 3.13 using version 2.60 of the `rstan` package. Raw data for all participants presented in the paper

and R code for running the models are available in a github repository at http://github.com/dyurovsky/XSIT-MIN. All three models—Statistical Accumulation, Single Referent, and Integrated—were fit to the data from Experiment 1 at the individual-participant level. Best-fitting parameters for Experiment 1 for each model were estimated by computing the mean value returned across 1000 samples.

While the Single Referent and Statistical Accumulation models capture some of the structure in the data in Experiment 1, 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). The one mismatch of the model to the data was in Switch trial performance for the 3- and 4-Referent conditions, in which the Integrated model predicted slightly lower performance than that actually exhibited by participants.

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 Referent and Statistical Accumulation models predict some of the variance in the new data, but leave much unexplained. The Integrated model makes near-perfect predictions about the new data—including the New Label condition—explaining 97% of the combined variance in the data from Experiments 1 and 2 (Table 3). Fig. 5 presents model predictions for all experimental data. Taken together, Experiments 1 and 2 and the Integrated model results thus provide strong evidence that learners track not only a single hypothesis for the most likely referent of a novel word, but also some approximation to distributional statistics; an approximation that becomes less precise as referential uncertainty increases.

4. 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 a synthetic 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 candidate referent other than the one that is currently favored; when there are few
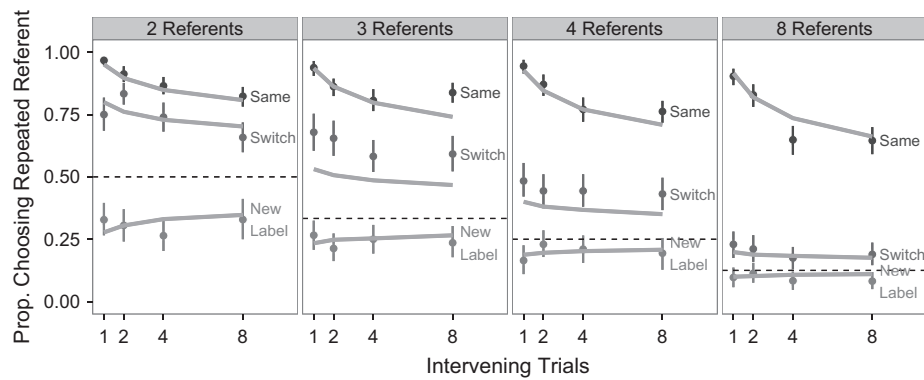
---

[4] This formulation is equivalent to using Luce's (Luce, 1959) Choice Axiom with the target having strength $M(L_o) + \frac{1-M(L_o)}{|O|}$ and each competitor having strength $\frac{1-M(L_o)}{|O|}$.

Fig. 5. Predictions of the Integrated model for all conditions in Experiment 1 and 2.

possibilities, 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.

This graded shift in representation 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; Vlach & Johnson, 2013; Yurovsky et al., 2014). This framework allowed us to estimate the effects of these constraints on learning to find the model that best fit the data—one 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 (Kachergis, Yu, & Shiffrin, 2013; Waxman & Gelman, 2009).

Our account also provides some insight into the conflicting results of previous experiments (Fig. 1). Because the amount of information participants encode about each non-hypothesized referent falls of proportionally to the number of referents presented, and because the amount of information participants remember falls of proportionally to the interval between successive exposures, statistical power to detect significant information falls of rapidly as the task grows in complexity. Thus, our prediction is that experiments that did not detect multiple-hypothesis tracking might well have failed because such effects would have been very small and would have required extremely large samples to distinguish from chance with any reliability. This pattern is further complicated by interactions in the way that participants encode and retrieve information as the same referents co-occur with multiple different words (Yurovsky et al., 2014; Yurovsky et al., 2013). Consequently, we believe that much of the previous confusion has arisen from a combination of measurement and statistical inference issues and a failure to appreciate the effects of particular task parameters on the expected effect size.

The shift from a computational to an algorithmic (or, psychological) 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, the literature contains a number of previous models of cross-situational learning aimed at fitting human-level performance in varying learning conditions with various instantiations of cognitive limitations (e.g., Fazly, Alishahi, & Stevenson, 2010; Kachergis et al., 2012;

Smith et al., 2011; Tilles & Fontanari, 2013; Yurovsky et al., 2014). Problematically, as demonstrated in a recent paper by Yu and Smith (2012), these seemingly distinct models can perfectly mimic each other at different parameter settings (see also, Townsend, 1990). These authors note that modeling choices peripheral to the central learning mechanism—e.g., attentional allocation, memory, choice rule—can be varied to produce many different patterns of learning.

Our goal in this paper was not to distinguish among these competing models, or to ultimately rule them out. Instead, our goal was to be agnostic as possible about the mechanisms underlying cognitive constraints and to ask instead how such constraints produce variation in the fidelity of mapping representations. To facilitate these inferences, we fit a large set of parametrically-varying data that imposes strong constraints on model parameters and modeling choices. In addition, we prevented overfitting by fixing model parameters using Experiment 1 and making parameter-independent predictions about learning that were supported in Experiment 2. This approach allowed us to gain insight about both the central learning mechanism and the constraining processes that together determine human performance.

Although cross-situational learning has been proposed as a potential acquisition mechanism for children (e.g. Pinker, 1989), the majority of experimental work has focused on adults. While children can learn from cross-situational evidence (Smith & Yu, 2008; Suanda et al., 2014; Vlach & Johnson, 2013), the mechanisms underlying these inferences could well be different from those operating in adults. Indeed, some recent findings suggest qualitative differences between children and adults, specifically in scenarios that require exclusion inferences (Ramscar, Dye, & Klein, 2013). Any inference from adult data to children's learning mechanisms remains necessarily speculative.

Nonetheless, as more developmental data become available, models like ours will be important tools in interpreting these data. Adults and children differ substantially in general cognitive abilities such as memory and attention (e.g. Gathercole, Pickering, Ambridge, & Wearing, 2004; Lane & Pearson, 1982). Our model suggests that even if there were continuity in learning *mechanisms* across age, the representations underlying cross-situational learning might still seem to shift between childhood and adulthood. For young children, even "simple" two-referent situations might be sufficiently challenging to prevent strong representation of multiple alternatives. Thus, interpretation of new data should be guided by predictions for memory- and attention-constrained learners.

We further note that connecting experimental data from children to the natural context of word learning may also require substantial work. Cross-situational learning experiments may impose additional cognitive demands on children (e.g., encoding many new words and unfamiliar objects) that are not representative of

the familiar circumstances in which children's word learning often takes place. In natural speech to children, referents are introduced into common ground and then discussed (Clark, 2003). In contrast, cross-situational tasks are intentionally stripped of the constellation of communicative, attentional, and linguistic cues that typically surround naming events (Frank, Tenenbaum, & Fernald, 2013; Gogate, 2010; Mintz, 2003), and each naming event appears in isolation, rather than being embedded in a coherent discourse (Frank et al., 2013; Rohde & Frank, 2014).

Further, while cross-situational learning tasks have typically studied the mapping process independent of the generalization process, children's representations of the meanings of even concrete nouns (e.g. *cup*) appear to follow an extended developmental trajectory, changing well into young-adulthood (Ameel, Malt, & Storms, 2008). This representational change is likely related to variability in the exemplars children are exposed to, variability in the contexts in which they are seen, and to the other words children have learned and the kinds of hypotheses they have entertained (Dautriche & Chemla, 2014; Hidaka & Smith, 2010). Thus, a full understanding of the processes of early word learning will necessarily require further analyses of the natural ecology of word learning and how it changes across development. Nonetheless, data and models of the kind presented here provide useful guiding principles for understanding word learning in the wild.

In sum, our work stands as a case study of how ideal learning models can inform psychological accounts of statistical learning. Although we focused on noun learning, our results are relevant for many problems in 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.

Appendix A

Experiment 1 showed that participants encode multiple referents in ambiguous naming situations, even under high levels of cognitive load. However, Experiment 1 leaves open the possibility that while participants encoded multiple referents, they did not map them to particular words. Experiment 2 was designed to rule out this unimodal familiarity account, showing that in the presence of a novel label, participants *dispreferred* the familiar object, suggesting that they encoded words, referents, and something about the relationship between them. However, these results are in-principle consistent with an alternative account in which participants track referent familiarity and word familiarity independently, and use a complex choice rule that selects the most familiar referent in the presence of a familiar word and the most unfamiliar referent in the presence of a novel word. Experiment 3 was designed to test this account.

In Experiments 1 and 2, the set of candidate referents for each word participants learned was distinct from the set of candidate referents for every other word. This experimental choice prevented participants from using knowledge about one word's referent to learn the correct referent of another novel word (c.f. Smith et al., 2011; Yurovsky et al., 2013). Choosing to conduct the experiment this way both allowed us to produce better estimates of learning fidelity across conditions and simplified the choice rule necessary for our cognitive model. In Experiment 3, we relax this constraint, however, allowing a more stringent test of the alternative account above. This time, the test set for each word contained both a previously co-occurring (and thus statistically correct) referent,

and a referent that had been seen more recently, but had co-occurred with a different word. In this way, Experiment 3 directly pitted statistical co-occurrence information against unimodal word and referent familiarity.

A.1. Method

Table A1 shows a design diagram comparing Experiment 3 to Experiments 1 and 2. In Experiment 3, participants again received two trials for each novel word. On the exposure trial (Ex-A), as before, they selected one of the candidate referents. On the test trial for each word (T-A), the set of candidate referents contained one that participants had previously seen on the Exposure trial for that word (a1), *and* one of the referents from the Exposure trial for the intervening word (b2). Our mapping account of the data in Experiments 1 and 2 predicts that participants should select the referent that co-occurred previously with the tested word (a1). In contrast, a familiarity account of the previous data predicts that participants should select the referent that co-occurred with the intervening word (b2).

As in the other experiments, we were interested in how learning scales both with the number of referents and with participants' guesses on Exposure trials. We thus tested two referent sizes between subjects: 2 and 4. We also tested all four possible combinations of guesses on Exposure trials for the two critical words within-subjects. That is, we compared the referent participants selected on Ex-A to the referent participants selected on Ex-B (Same vs. Same), the referent participants selected on Ex-A to a referent they did not select on Ex-B (Same vs. Switch), etc.

A.1.1. Participants

As in Experiments 1 and 2, participants for Experiment 3 were recruited from Amazon Mechanical Turk under the constraint that they had a US IP address. Each HIT paid 30 cents for completion. Because Experiment 3 contained fewer trials per within-subjects condition than Experiments 1 and 2, we posted 250 HITs for each Referent condition for a total of 500 paid HITS. Participants were again paid for multiple HITs, but only data from their first was included in the final set (excluded 8 HITs) and an additional HIT

Table A1
Schematic of the design for Experiment 3. In contrast to Experiments 1 and 2, the test trial (T-A) for each word pitted a previously co-occurring referent (a1) against a more recently exposed familiar competitor referent (b2).

Trial	Experiment 1			Experiment 2			Experiment 3		
	Word	Ref 1	Ref 2	Word	Ref 1	Ref 2	Word	Ref 1	Ref 2
Ex-A	A	a1	a2	A	a1	a2	A	a1	a2
Ex-B	B	b1	b2	B	b1	b2	B	b1	b2
T-A	A	a1	a3	C	a1	a3	A	a1	b2

Table A2
Predictor estimates with standard errors and significance information for a logistic mixed-effects model predicting word learning in Experiment 1. The model was specified as `Correct ~ Log(Referents) * Target + Log(Referents) * Competitor + offset(logit(1/`*Referents*`)) + (TrialType | subject)`.

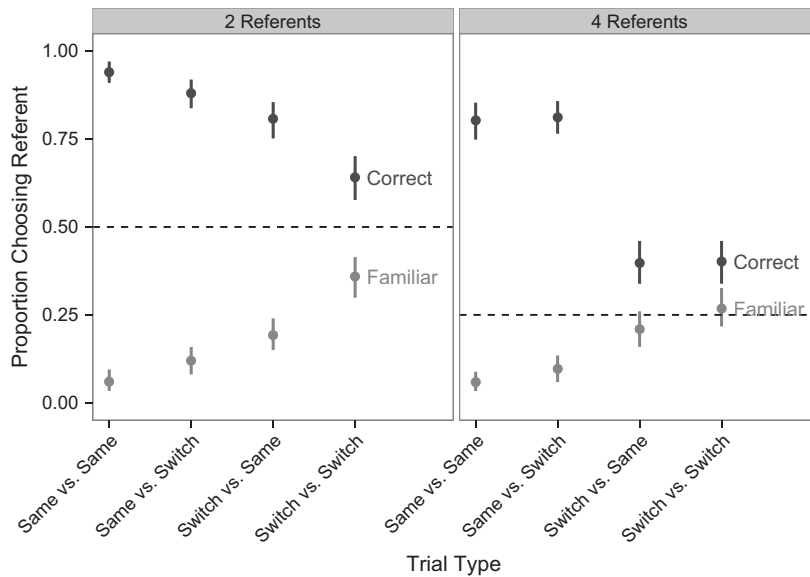| Predictor	Estimate	Std. error	z value	p value	
Intercept	3.27	0.45	7.27	<.001	***
Log(Referents)	−0.48	0.37	−1.32	0.19	
Switch target	−0.92	0.42	−2.17	0.03	*
Switch competitor	−1.77	0.40	−4.40	<.001	***
Log(Referents) * Switch target	−0.75	0.36	−2.07	0.04	*
Log(Referents) * Switch competitor	1.30	0.34	3.76	<.001	***

Fig. A1. Proportion of participants choosing the Correct (co-occurring) and Familiar (most recently seen) referent for each of the four trial types in Experiment 3. Each datapoint represents ~230 participants. Error bars indicate 95% confidence intervals computed by non-parametric bootstrap. In all cases, participants chose the Correct referent at above chance levels, and also significantly more often than the Familiar referent, indicating that they learned a mapping between words and objects even on Switch trials. Performance decreased predictably when the number of referents increased and when the Correct referent was the one not selected by participants on Exposure trials (*Switch vs. X*).

was posted. In addition, data was again excluded from the final sample if participants did not give correct answers for familiar trials (35 HITs). The final sample thus comprised 465 unique participants, 234 in the 2 Referents condition and 231 in the 4 Referents condition.

A.1.2. Stimuli, design, and procedure

Experiment 3 used the same Stimuli and general trial structure as Experiments 1 and 2 (Table A1). Each participant received either the 2 Referents or the 4 Referents condition. Each participant was exposed to 8 words, 4 of which were tested and 4 of which provided the Familiar competitors. One of the words was tested in each possible Target vs. Competitor condition (Same v. Same, Same vs. Switch, Switch vs. Same, Switch vs. Switch). In all cases, one Competitor Exposure trial (Ex-B) occurred between the Exposure and Test trials for the Target word, as in the Interval 2 conditions of Experiments 1 and 2.

A.2. Results

Do learners encode mappings for multiple candidate referents, or do they instead encode words and referents independently? If participants encoded mappings, we would expect them to select the Correct referent more frequently than expected by chance on test trials, even when the Correct referent was one they did not previously select on its Exposure trial. To test this prediction, we again fit adjusted logistic regressions to each condition (Correct ∼ 1 + offset(logit(1/Referents))). For both 2 and 4 Referents, for all test types, participants selected the Correct referent at levels higher than predicted by chance (smallest $\beta = .58, z = 4.25$, all *ps* <.001). In the 2 Referent condition, this necessarily means that they chose the Correct referent more frequently than the Familiar referent, as these were the only two options at test. However, in the 4 Referent condition, two other novel competitors were available. To show that participants distinguished between the Correct and Familiar referents in this condition, we need to further show that they were chosen at different rates We thus fit an adjusted logistic regression to determine

whether Referent type was a significant predictor of performance (Correct ∼ 1 + Type + offset(logit(1/Referents))). Indeed, for all test types, the Correct referent was chosen more frequently than the Familiar referent (smallest $\beta = .61, z = 3.09$, all *ps* <.01). Thus, learners encoded more multiple word-referent mappings, and not just multiple words and referents (Fig. A1).

To examine differences across the conditions we tested, we fit a mixed-effects model to determine how performance varied with test type and number of Referents (Table A2). This model showed a significant Intercept term, indicating globally above-chance selection of Correct referents. It further showed significant main effects Target Type, and Competitor Type, indicating that participants performed best when both the Target and Competitor were referents they had previously hypothesized on Exposure trials. Finally, the model showed significant interactions between both Target and Competitor types and number of Referents, indicating that at 4 Referents the identity of the Target referent had more effect than the identity of the Competitor referent.[5]

In sum, the results of Experiment 3 provide strong support for the account of Experiments 1 and 2 given in the main text: when encountering a novel word, participants in these experiments encoded both their hypothesized referent, and multiple additional referents at above chance levels. Even when the correct referent at test was one that participants had seen but not hypothesized (as in Switch trials in Experiment 1), and even when one of the competitors was seen more recently and thus more familiar, participants selected the correct referent at above chance levels. Further, as in the previous Experiments, performance tracked predictably with both the number of Referents in training and whether the Target

---

[5] Although participants did not receive feedback, one might nonetheless be concerned that because the Correct referent in this task was always the referent that was seen two trials ago, participants' above-chance performance on this task might have been due to learning a meta-strategy of selecting the referent from two trials ago. Such an account would predict increased performance over the course of the experiment as participants discovered this strategy. To test this alternative, we added an additional term to the mixed-effects model: Trial number. This regression showed a statistically significant decrease in performance over the course of the experiment (smallest $\beta = -.132, z = -4.93, p < .001$), ruling out this account.

(and Competitor) referent was previously selected by participants on Exposure trials. These results thus provide further support for an Integrated model of cross-situational learning in which people encode both a strong single hypothesis and weaker but reliable distributional information about alternative candidate referents.

References

Ameel, E., Malt, B., & Storms, G. (2008). Object naming and later lexical development: From baby bottle to beer bottle. *Journal of Memory and Language, 58*, 262–285.

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*, 390–412.

Blythe, R. A., Smith, K., & Smith, A. D. M. (2010). Learning times for large lexicons through cross-situational learning. *Cognitive Science, 34*, 620–642.

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

Clark, E. V. (2003). *First language acquisition*. Cambridge University Press.

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*, e57410.

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

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*, 406–419.

Fazly, A., Alishahi, A., & Stevenson, S. (2010). A probabilistic computational model of cross-situational word learning. *Cognitive Science, 34*, 1017–1063.

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.

Frank, M. C., Tenenbaum, J. B., & Fernald, A. (2013). Social and discourse contributions to the determination of reference in cross-situational word learning. *Language Learning and Development, 9*, 1–24.

Gallistel, C. R., Fairhurst, S., & Balsam, P. (2004). The learning curve: Implications of a quantitative analysis. *Proceedings of the National Academy of Sciences, 101*, 13124–13131.

Gathercole, S. E., Pickering, S. J., Ambridge, B., & Wearing, H. (2004). The structure of working memory from 4 to 15 years of age. *Developmental Psychology, 40*, 177.

Gogate, L. J. (2010). Learning of syllable–object relations by preverbal infants: The role of temporal synchrony and syllable distinctiveness. *Journal of Experimental Child Psychology, 105*, 178–197.

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

Hidaka, S., & Smith, L. B. (2010). A single word in a population of words. *Language Learning and Development, 6*, 206–222.

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*, 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*, 133–142.

Lane, D. M., & Pearson, D. A. (1982). The development of selective attention. *Merrill-Palmer Quarterly, 28*, 317–337.

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

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*, 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.

Mintz, T. H. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition, 90*, 91–117.

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.

Ramscar, M., Dye, M., & Klein, J. (2013). Children value informativity over logic in word learning. *Psychological Science, 24*, 1017–1023.

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

Rohde, H., & Frank, M. C. (2014). Markers of topical discourse in child-directed speech. *Cognitive Science, 38*, 1634–1661.

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

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

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.

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.

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*, 174–215.

Suanda, S. H., Mugwanya, N., & Namy, L. L. (2014). Cross-situational statistical word learning in young children. *Journal of Experimental Child Psychology, 126*, 395–411.

Tilles, P. F., & Fontanari, J. F. (2013). Reinforcement and inference in cross-situational word learning. *Frontiers in Behavioral Neuroscience, 7*.

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.

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*, 375–382.

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

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. *The Journal of Neuroscience, 32*, 11228–11240.

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., 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.

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.