A large-scale comparison of cross-situational word learning models

Anonymous CogSci submission

Abstract

How people learn word meanings from ambiguous situations has been a frequent target of empirical work as well as numerous computational models, and yet there is little consensus on the main mechanisms underpinning cross- situational word learning. The lack of consensus is in part due to the large number of disparate studies and models, and lack of systematic model comparisons across a wide range of studies. This study seeks to remedy this, comparing the performance of several extant cross-situational word learning models on a dataset consisting of 1,696 participants spread across 44 experimental conditions. Using cross-validated parameters, we fit multiple models representing theories of both associative learning and hypothesis-testing theories of word learning, and discuss which models mimic each other, and which mechanisms are overall best able to account for the data.

Keywords: cross-situational word learning; model comparison; language learning

Introduction

Learning word meanings is a crucial aspect of the process of language acquisition. If a learner's task is to connect the words they hear to the potential referents and eventually to generalizable meanings, the challenge appears formidable (Quine, 1960): words are heard in rich and cluttered contexts. Yet children learn words very quickly. How do they do this?

An array of social (Yurovsky & Frank, 2017), pragmatic (Clark, 1987; Markman, 1992), and linguistic (Gleitman, 1990) cues are often available to reduce a learner's uncertainty. Yet at its heart, these information sources can be seen as reducing the uncertainty inherent in the task of *cross-situational mapping* (Frank, Goodman, & Tenenbaum, 2009): that is, cross-referencing which referents a word has consistently appeared with across multiple uses (Carey & Bartlett, 1978; Gleitman, 1990). If learners are able to learn cross-situationally, then they might be able to pool information about reference (and hence about word meaning) over time and learn even in the absence of perfectly disambiguated input in any particular situation. Can children (or even adults) learn in this way?

The general idea of cross-situational word learning has been adopted into a variety of experimental paradigms for testing the word learning ability of infants (e.g., Smith & Yu, 2008; Vlach & Johnson, 2013), children (e.g., Akhtar & Montague, 1999; Suanda & Namy, 2012; Vlach & Sandhofer, 2012), and adults (e.g., Yu & Smith, 2007; Smith, Smith, & Blythe, 2011). In such experiments, participants are typically presented with a series of training trials comprised of multiple possible referents (e.g., an array of 2-4 novel objects) and one or more spoken nonce words, and (when possible) instructed to learn the meaning(s) of the presented words. These training trials typically present 10–20 distinct word-referent pairs to be learned across 10–50 trials, presenting each word/referent 3–10 times across a total span of < 10 minutes. Learning is then typically assessed by presenting each word and asking learners to make a forced choice of the best referent, either from the entire set or from a small subset of referents.

People of all ages are able to learn word-object mappings in such paradigms at a level significantly above chance, but their performance is typically far from perfect in all but the most trivial versions of the task (Yurovsky & Frank, 2017). Numerous training manipulations have been found to influence performance. How often mappings are presented (Kachergis, Yu, & Shiffrin, 2016; Suanda & Namy, 2012), the interval between appearances of a mapping (Vlach & Johnson, 2013; Vlach & Sandhofer, 2012), and the number of mappings presented per trial (Yu, 2008) all have been shown to affect learning performance. In sum, the evidence is strong that people can learn word cross-situationally: we are somehow able to integrate evidence (co-occurrence of words and objects) across trials and use that evidence to learn and remember new word meanings in cross-situational learning experiments. But what are the psychological mechanisms supporting this process? And critically, do these mechanisms support naturalistic word learning?

Computational models provide one strategy for instantiating proposals about mechanism. Such models can then be applied to naturalistic data to make a first assessment of whether cross-situational learning could be a viable strategy for word learning "in the wild". Accordingly, there has been great interest in creating and testing computational models of this task. These models often represent distinct intuitions of how learning operates, variously representing the learning problem as logical inferencing (e.g., Siskind, 1996), proposing and testing hypotheses (e.g., Trueswell, Medina, Hafri, & Gleitman, 2013), or accumulating graded associations (e.g., Fazly, Alishahi, & Stevenson, 2010; Kachergis, Yu, & Shiffrin, 2012; Tilles & Fontanari, 2012). Extant cross-situational word learning models are too numerous to review exhaustively here.

It is unknown which of these models – or even which broad

class of models – best fits the breadth of experimental data. While most models of cross-situational learning have typically been tested against at least a few datasets, to our knowledge there has been no systematic comparison of models across a wide variety of experimental conditions. Identifying the model or models that best predict human performance is thus an important goal for consolidating this broad literature. More broadly, identifying such models could be an important step in assessing the contribution of cross-situational learning to word learning more generally.

Thus, the goal of our current study is to test the ability of a range of cross-situational word learning models in accounting for a wide range of experimental data. Our goal is to scale up both the number of models evaluated and the number of experiment designs (and participants) being used for evaluation. Different models often contain a range of parameters that must be fit to data, an issue that has hindered previous comparison work. Thus, we focus on out-of-sample prediction of human data using cross-validated parameter fits. Finally, we will interpret the results, and what they entail for theories of cross-situational word learning.

Method

Procedure

In a nutshell, our goal is to fit a set of models to a set of datasets; members of each of these sets are listed below. Each model typically is defined as a function that takes input data and a group of parameter values (often between 2–4) and then can be evaluated on test trials (typically multi-alternative forced choice test). Each model then returns predictions on each test trial, which can be scored against the "correct" (intended by the design) answer and averaged to produce an analogue to human performance. Our primary outcome measure is the correlation between each model's performance for each item in a specific experimental condition and item-level human performance on that same experimental condition.

Although input data and test trials are features of specific experimental tests, each model has parameters that must be specified. Typically the settings of these parameters are critical to the predictive performance of the model. Hence, to ensure a fair comparison between models, these parameters must be fit to data. In all cases, we optimize parameter values to minimize the sum of squared error (SSE) between model and human performance, weighting the contribution of each condition's SSE by the number of participants in that condition. We optimize parameters using differential evolution via the DEoptim R package (Mullen, Ardia, Gil, Windover, & Cline, 2011).

We conducted two evaluations of models. First, we use cross-validation across conditions to reduce overfitting. We optimize each model's parameters by leaving out data from 20% of conditions to test generalization via optimization on the initial 80% (5-fold cross-validation). Second, in the *group fit* regime we fit each model to all conditions and datasets to produce a single best-fitting parameter set per model. We then evaluated the correlation between these fitted models and

the same data they were fit to. This evaluation thus makes maximal use of our data but incorporates some amount of overfitting. However, it allows us to make the best test of generalization to a selection of other experiments from the literature for which we do not have item-level accuracy.

Models

We fit the following associative models: Fazly et al. (2010), Kachergis et al. (2012), three simplified variations of that model, a Bayesian decay model, and a baseline co-occurrence counting model. We also fit two hypothesis-testing models: guess-and-test (Blythe, Smith, & Smith, 2010; Medina, Snedeker, Trueswell, & Gleitman, 2011) and propose-but-verify (Trueswell et al., 2013). Finally, we fit two hybrid models, that store multiple, graded hypotheses–but only a subset of all possible associations: pursuit (Stevens, Gleitman, Trueswell, & Yang, 2017), and a stochastic version of the Kachergis model. These models and their free parameters are briefly described below.

Baseline Co-occurrence Model   The baseline model simply tallies the co-occurrences of any words and objects appearing together on each trial, accumulating the counts in a word × object matrix.

Fazly et al. Model   The Fazly et al. (2010) model represents the meaning of each word w as a probability distribution $p(\cdot|w)$ over the objects appearing across the trials, and incrementally learns these distributions trial-by-trial. The association between a given w and o grows more quickly if $p(w|o)$ is already high (a familiarity bias), unless some other word has a strong association with o; thus, associations are competitive. A parameter λ is a small smoothing constant added to both the numerator and denominator of each $p(w|o)$, further adjusted in the denominator by a β parameter representing the upper bound on the number of symbol types.

Kachergis et al. Model   The Kachergis et al. (2012) associative model assumes that learners associate all presented words with all visible objects to some extent, but that they selectively attend more to some of the presented word-object pairings. Specifically, greater attention is directed to pairings that have previously co-occurred (a familiarity bias), but is also directed to stimuli that have no strong associates, tracked via the entropy (H) of their association strengths (e.g., novel stimuli, or stimuli that have diffuse associations with several stimuli). Thus, the model grows its association matrix as it experiences each trial, dynamically apportioning a fixed amount of associative weight (learning rate χ) among the possible word-object associations, with the relative weight of the familiarity bias and the uncertainty bias determined by parameter λ. Finally, all association strengths decay from trial to trial at a rate controlled by parameter α.

The update rule for adjusting the association $M_{w,o}$ between a given word w and object o on a given trial is:

$$M_{w,o} = \alpha M_{w,o} + \frac{\chi \cdot e^{\lambda \cdot (H(w)+H(o))} \cdot M_{w,o}}{\sum_{w \in W} \sum_{o \in O} e^{\lambda \cdot (H(w)+H(o))} \cdot M_{w,o}} \quad (1)$$

At test, given word *w* learners choose referent *o* from the *m* given alternatives in proportion to associative strength $M_{w,o}$ (i.e. Luce's choice rule).

We also fit a stochastic *sampling* version of this model (see Kachergis & Yu, 2017), which uses the same update equation, but instead of updating all possible word-object associations samples only a single presented object for each word on a trial. This sampling version of the Kachergis model can still build multiple graded associations for each word (or object), but on any given run will accumulate only a sparse, randomized version of what the full associative model would build.

Strength-, Uncertainty-, and Novelty-Biased Models We also test three variations on the Kachergis et al. model that use only a subset of the mechanisms of the full model in order to understand the contributions of each mechanism. The *strength model* lacks the uncertainty bias term, and thus only implements a bias for familiar (i.e., already-strong) associations, with learning rate (χ) and decay α parameters.

The *uncertainty model* lacks the strength bias term, and thus only implements a bias for stimuli with uncertain (high-entropy) associations. The *novelty model* lacks the strength bias term, and substitutes novelty for the entropy terms (e.g., 1/(frequency(*w*)+1) instead of $H(w)$). These models still have all three parameters of the original Kachergis model.

Bayesian Decay Model This previously-unpublished model updates the $p(w|o)$ and the joint probability $p(w,o)$ from trial to trial according to a likelihood function that reinforces the association between *w* and *o* when they co-occur (scaled by parameter δ), and penalizes all associations that are not occurring on the trial (scaled by parameter α). Thus, in contrast to other incremental associative learning models considered here (e.g., Fazly and the Kachergis class of models), this model globally updates the entire association matrix on each trial – both co-occurring pairs and non-co-occurring pairs, and re-normalizes the tracked probabilities at each time step. At test, this model uses a softmax choice rule with parameter τ.

Rescorla-Wagner Model This associative model is inspired by the prediction-error-based learning model of Rescorla & Wagner (1972). Objects on each trial serve as cues to predict which words will be heard. When a given word is heard, the associations between that word and each object on the trial are scaled at a learning rate β in proportion to the magnitude of the difference between the prediction and the maximum association value (λ). Each trial, the association matrix is subject to decay (parameter α).

Propose-but-verify Model In the propose-but-verify hypothesis testing model (Trueswell et al., 2013), one of the presented referents is selected at random for any word heard that has no remembered referent. The next time that word occurs, the previously-proposed referent is remembered with probability α, a free parameter. If the remembered referent is verified to be present, the future probability of recalling

the word is increased by ε (another free parameter). If the remembered referent is not present, the old hypothesis is assumed to be forgotten and a new proposal is selected from the available referents. This model implements trial-level mutual exclusivity by selecting new proposals only from among the referents that are not yet part of a hypothesis.

Guess-and-test Model The guess-and-test hypothesis testing model is based on the description given by Medina et al. (2011) of a one-shot (i.e. "fast mapping") learning, which posits that "i) learners hypothesize a single meaning based on their first encounter with a word, ii) learners neither weight nor even store back-up alternative meanings, and iii) on later encounters, learners attempt to retrieve this hypothesis from memory and test it against a new context, updating it only if it is disconfirmed." In summary, guess-and-test learners do *not* reach a final hypothesis by comparing multiple episodic memories of prior contexts or multiple semantic hypotheses. We give this model two free parameters: a probability of successful encoding (*s*, hypothesis formation), and a probability *f* of forgetting a hypothesis at retrieval. This model is quite similar to the guess-and-test model formally analyzed by Blythe et al. (2010) to determine the theoretical long-term efficacy of cross-situational word learning.

Pursuit Model The pursuit model (Stevens et al., 2017) represents a hybrid approach, potentially storing graded associations involving a given word and multiple referents (or vice-versa), while on a given trial greedily pursuing the single strongest association for each presented word. If the strongest associated referent for a given word *w* is present on the trial, the association is strengthened at a rate determined by γ. If the strongest referent for *w* is not present, the association is weakened (scaled down by $(1-\gamma)$) and the association with a random other available referent is strengthened. Novel words are given an initial association of strength γ with the available referent whose strongest association is the smallest, implementing an uncertainty bias in forming associations for novel words. A given word-object association only enters the lexicon if $p(o|w) > \theta$, a threshold parameter.

Data

The modeled data are average accuracies from 726 word-object pairs in 44 experimental conditions, in which a total of 1696 subjects participated. Table 3 shows a variety of characteristics of each experimental condition, including the number of trials (range: 18-108), the number of words and referents presented per training trial (1-4), the number of total to-be-learned pairs (12-24), the number of participants per condition, and the average accuracy and SD of performance. Most of these data have been previously published: data from 13 of the conditions were reported in Kachergis & Yu (2013), data from 12 conditions are from Kachergis et al. (2012), and data from another nine conditions are from Kachergis et al. (2016). However, data from several conditions run in Chen Yu's lab at Indiana University have never before been published. Each condition consists of an ordered list of training

trials consisting of 1-4 words and 2-4 objects per trial.

Results

Table 1 shows the sum of squared error (SSE) and correlation (*r*) for each model's best-fitting parameters vs. average human performance on the 726 items, both for the CV fits and for the all-condition fits, along with the number of fitted parameters in each model (P). The results of the two fitting procedures yielded fairly similar SSEs and correlations, and a consistent rank-ordering of the models, with the Kachergis sampling model performing best, followed closely by the Fazly et al. model and then the associative Kachergis et al. model. The hypothesis testing models (Propose-but-Verify and Guess-and-Test) and the Pursuit model fit less well than the 0-parameter baseline co-occurrence counting model (not in the table), which had SSE = 32.0 and $r = 0.53$. Correlations between data and models are presented in Figure 1.

How related to each other are these models? Table 2 shows the correlations between model predictions made using the group-optimized parameters, with the names of the models abbreviated for space (*Ks*: Kachergis sampling, *Fa*zly, *K*achergis, *N*ovelty, *U*ncertainty, *Ba*yesian decay, *R*escorla-Wagner, *S*trength, Propose-but-*V*erify, *G*uess-and-Test, *P*ursuit). These results are discussed below.

Model	CV SSE	r	All SSE	r	P
Kachergis samp	17.74	0.74	17.76	0.74	3
Fazly	18.72	0.73	18.77	0.73	2
Kachergis	18.92	0.72	19.07	0.72	3
Novelty	22.02	0.67	22.01	0.67	2
Uncertainty	23.06	0.69	23.02	0.69	3
Bayes decay	23.95	0.62	23.71	0.63	3
Rescorla-Wagner	31.80	0.62	31.79	0.62	3
Strength	39.05	0.46	39.07	0.46	2
Propose-Verify	42.92	0.35	31.30	0.50	2
Guess-Test	43.01	0.32	31.92	0.48	2
Pursuit	49.33	0.31	46.75	0.35	3

Table 1: Cross-validated and group model fits.

	Ks	F	K	N	U	B	R	S	V	G
F	.87	-	-	-	-	-	-	-	-	-
K	.92	.84	-	-	-	-	-	-	-	-
N	.87	.79	.87	-	-	-	-	-	-	-
U	.82	.81	.83	.74	-	-	-	-	-	-
B	.75	.79	.79	.82	.64	-	-	-	-	-
R	.51	.58	.50	.50	.47	.56	-	-	-	-
S	.83	.84	.78	.79	.80	.70	.67	-	-	-
V	.47	.54	.45	.45	.44	.51	.95	.64	-	-
G	.60	.59	.60	.74	.53	.67	.53	.65	.45	-
P	.41	.47	.39	.43	.35	.52	.58	.49	.53	.58

Table 2: Correlations of models' predictions.

Generalization Experiment Results

Using the parameters optimized to each model for all of the conditions, we simulate model performance for a selection of published experiments and compare to participants' group-level performance for these conditions.

Discussion

We set out to conduct a systematic comparison of several models of cross-situational word learning models across a wide variety of experimental conditions. Using a database of 1696 participants in 44 conditions, we found best-fitting parameters for 11 different models under two different regimes: 1) simultaneously fitting all of the data, and 2) 5-fold cross-validation, testing generalization to the final 20% of conditions after training on 80% of them. The results of both fitting regimes were consistent and clear: shown in Table 1, the associative models achieved much better fits (lower SSE and higher correlation) than the hypothesis-testing models (Guess-and-Test; Medina et al. (2011) and Propose-but-Verify; Trueswell et al. (2013)), which made strongly-correlated predictions (see Table 2: .95). The hybrid Pursuit model (Stevens et al., 2017) fared even more poorly, in part because it often outperformed humans ($M_{mod} = 0.58$), $M_{hum} = 0.48$).

Among the associative models, the Fazly et al. (2010) model and the Kachergis et al. (2012) model both fare quite well, and the sampling version of the Kachergis et al. model performs slightly better, albeit with 3 free parameters compared to 2 in the Fazly et al. model. The predictions of the top three models are more correlated with each other than with the human data (see Table 2): Fazly vs. Kachergis sampling: .87, Fazly vs. Kachergis: .84, Kachergis vs. sampling: .92. But the Fazly model also showed a .84 correlation with the Rescorla-Wagner model, and both Kachergis models showed a .87 correlation with the simpler novelty-biased associative model–stronger than for the uncertainty-bias mechanism used by these models.

The Kachergis and Fazly models make maximally divergent predictions in the conditions with fewer words than objects per trial (esp. 1x3, 1x3 6/18 4AFC, and 1x4), which hints at a difference in mechanism for these models and may guide the design of a new experiment to diagnose which is better able to account for human behavior.

In summary, we have shown not only how a large model comparison can help rank models from most to least able to account for human data, but also how such comparisons can identify models that mimic each other, as well as identifying experimental designs that might aid in distinguishing models. Future work should extend this comparison to models and datasets beyond those tested here, and in particular to developmental experiments in order to test the models' generality for explaining word learning across a lifetime.
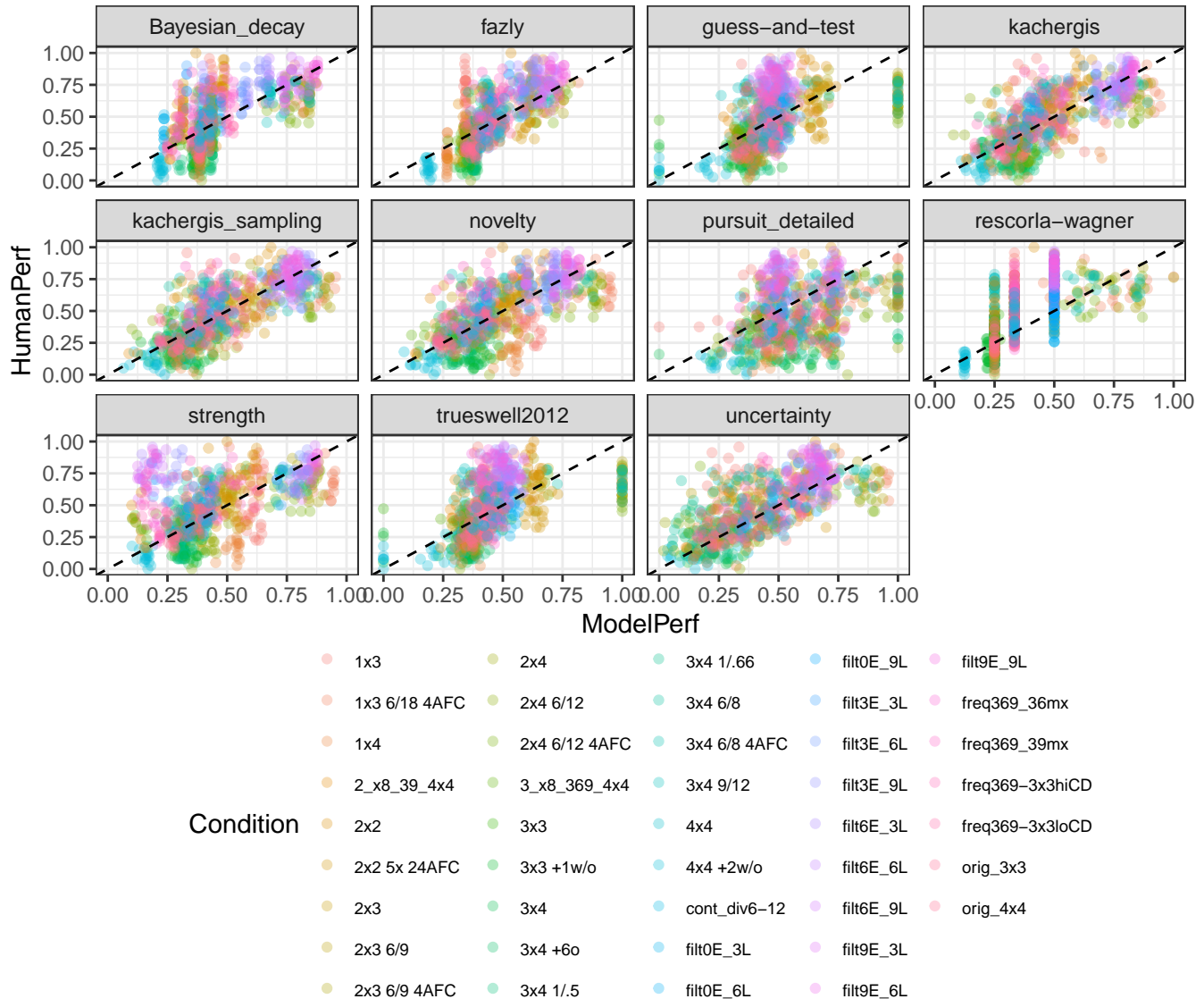
Figure 1: Human vs. model performance with a single best-fitting set of parameters for all conditions.

References

Akhtar, N., & Montague, L. (1999). Early lexical acquisition: The role of cross–situational learning. *First Language*, *19*, 34–358.

Blythe, R. A., Smith, K., & Smith, A. D. M. (2010). Learning times for large lexicons through cross-situational learning. *Cognitive Science*, *34*(4), 620–642.

Carey, S., & Bartlett, E. (1978). Acquiring a single new word. *Papers and Report on Child Language Development*, *15*, 17–29.

Clark, E. V. (1987). The principle of contrast: A constraint on language acquisition. In B. MacWhinney (Ed.), *Mechanisms of language aquisition*. Hillsdale, NJ: Erlbaum.

Fazly, A., Alishahi, A., & Stevenson, S. (2010). A Probabilistic Computational Model of Cross-Situational Word Learning. *Cognitive Science*, *34*(6), 1017–1063.

Frank, M. C., Goodman, N. D., & Tenenbaum, J. B. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psych. Science*, *20*(5), 578–585.

Gleitman, L. (1990). The structural sources of word meaning. *Language Acquisition*, *1*, 3–55.

Kachergis, G., & Yu, C. (2013). More naturalistic cross-situational word learning. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th annual conference of the cognitive science society*. Austin, TX.

Kachergis, G., & Yu, C. (2017). Observing and modeling developing knowledge and uncertainty during cross-situational word learning. *IEEE Transactions on Cognitive and Developmental Systems*.

Kachergis, G., Yu, C., & Shiffrin, R. M. (2012). An

associative model of adaptive inference for learning word–referent mappings. *Psychonomic Bulletin and Review*, *19*(2), 317–324.

Kachergis, G., Yu, C., & Shiffrin, R. M. (2016). A bootstrapping model of frequency and contextual diversity effects in word learning. *Cognitive Science*. `http://doi.org/10.1111/cogs.12353`

Markman, E. M. (1992). Constraints on word learning: Speculations about their nature, origins and domain specificity. In M. R. Gunnar & M. P. Maratsos (Eds.), *Modularity and constraints in language and cognition: The minnesota symposium on child psychology* (pp. 59–101). Hillsdale, NJ: Erlbaum.

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

Mullen, K., Ardia, D., Gil, D., Windover, D., & Cline, J. (2011). DEoptim: An R package for global optimization by differential evolution. *J. of Statistical Software*, *40*(6), 1–26.

Quine, W. V. O. (1960). *Word and object*. Cambridge, MA: MIT Press.

Rescorla, R., & Wagner, A. (1972). A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In *Conditioning ii: Current research and theory*.

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*(3), 480–498.

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

Stevens, J. S., Gleitman, L. R., Trueswell, J. C., & Yang, C. (2017). The pursuit of word meanings. *Cognitive Science*, *41*, 638–676.

Suanda, S. H., & Namy, L. L. (2012). Detailed behavioral analysis as a window into cross-situational word learning. *Cognitive Science*, *36*(3), 545–559.

Tilles, P. F., & Fontanari, J. F. (2012). Minimal model of associative learning for cross-situational lexicon acquisition. *Journal of Mathematical Psychology*, *56*(6), 396–403.

Trueswell, J. C., Medina, T. N., Hafri, A., & Gleitman, L. R. (2013). Propose but verify: Fast mapping meets cross-situational word learning. *Cognitive Psychology*, *66*(1), 126–156.

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

Vlach, H. A., & Sandhofer, C. M. (2012). Fast mapping across time: Memory processes support children's retention of learned words. *Frontiers in Developmental Psychology*, *3*(46), 1–8.

Yu, C. (2008). A statistical associative account of vocabulary growth in early word learning. *Language Learning and Development*, *4*(1), 32–62.

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

Yurovsky, D., & Frank, M. C. (2017). Beyond naïve cue combination: Salience and social cues in early word learning. *Developmental Science*, *20*(2), e12349.

Name	Trials	W/Tr	O/Tr	Items	Acc	SD	N
3x4	36	3	4	18	0.19	0.12	25
3x4 1/.5	36	3	4	18	0.22	0.11	25
3x4 1/.66	36	3	4	18	0.21	0.09	25
3x4 +6o	36	3	4	18	0.27	0.12	20
2x4	54	2	4	18	0.30	0.15	33
3x3 +1w/o	54	3	3	18	0.17	0.08	39
4x4 +2w/o	54	4	4	18	0.10	0.05	39
1x3 6/18 4AFC	108	1	3	18	0.67	0.10	43
2x3 6/9 4AFC	54	2	3	18	0.69	0.11	38
2x4 6/12 4AFC	54	2	4	18	0.62	0.12	31
3x4 6/8 4AFC	36	3	4	18	0.69	0.09	36
1x3	108	1	3	18	0.74	0.11	23
2x3	54	2	3	18	0.58	0.07	23
2x4 6/12	54	2	4	18	0.33	0.18	14
3x4 6/8	36	3	4	18	0.42	0.14	13
2x3 6/9	54	2	3	18	0.55	0.11	32
3x4 9/12	54	3	4	18	0.69	0.08	33
1x4	108	1	4	18	0.19	0.10	40
1x3	108	1	3	18	0.39	0.12	40
4x4	27	4	4	18	0.31	0.07	77
3x3	36	3	3	18	0.43	0.08	36
2x2	54	2	2	18	0.79	0.11	19
2x2 5x 24AFC	60	2	2	24	0.51	0.12	46
filt0E_3L	18	2	2	12	0.38	0.09	31
filt3E_3L	27	2	2	12	0.72	0.12	30
filt6E_3L	36	2	2	12	0.71	0.09	30
filt9E_3L	45	2	2	12	0.70	0.08	31
filt0E_6L	36	2	2	12	0.47	0.11	27
filt3E_6L	45	2	2	12	0.67	0.11	27
filt6E_6L	54	2	2	12	0.79	0.09	27
filt9E_6L	63	2	2	12	0.75	0.09	27
filt0E_9L	54	2	2	12	0.54	0.09	31
filt3E_9L	63	2	2	12	0.83	0.08	31
filt6E_9L	72	2	2	12	0.82	0.09	31
filt9E_9L	81	2	2	12	0.86	0.06	31
2_x8_39_4x4	27	4	4	18	0.41	0.16	30
3_x8_369_4x4	27	4	4	18	0.33	0.06	74
freq369-3x3loCD	36	3	3	18	0.33	0.08	102
freq369-3x3hiCD	36	3	3	18	0.56	0.12	26
freq369_36mx	36	3	3	18	0.45	0.16	62
freq369_39mx	36	3	3	18	0.62	0.14	66
orig_4x4	27	4	4	18	0.27	0.07	88
orig_3x3	36	3	3	18	0.43	0.08	104
cont_div6-12	36	3	3	18	0.43	0.11	40

Table 3: Summary of modeled datasets.