More Naturalistic Cross-situational Word Learning

George Kachergis[1] and Chen Yu[2]
[1]george.kachergis@gmail.com, [2]chenyu@indiana.edu
[1]Psychology Department, Leiden University, the Netherlands
[2]Department of Psychological & Brain Science / Cognitive Science Program, Indiana University, USA

Abstract

Previous research has found that people can use word-object co-occurrences from ambiguous situations to learn word meanings (e.g., Yu & Smith, 2007). However, most studies of cross-situational learning present an equal number of words and objects, which may simplify the problem by, for example, encouraging learners to use assumptions such as each word going with one object. This paper presents several conditions in which the number of words and objects do not match: either additional objects appear at random, or objects appear sometimes without their intended words. These manipulations do generally hurt learning in comparison to balanced conditions, but people still learn a significant proportion of word-object pairings. The results are explored in terms of statistics of the training trials—including contextual diversity and context familiarity—and with the uncertainty- and familiarity-biased associative model. Parametric differences between conditions hint at hidden cognitive constructs.

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

Introduction

Human infants learn words quite quickly despite many challenges facing them, including uncertainty and ambiguity in the language environment. Recent research has studied how learners may acquire word meanings from regularities in the co-occurrence of words and referents (e.g., objects). Such cross-situational statistical word learning relies on two assumptions: 1) that spoken words are often relevant to the visible environment, and 2) that learners can to some extent remember the co-occurrence of multiple words and objects in a scene. Thus, as words and their intended referents are observed in different situations over time, learners can apprehend the correct word-object mappings. Relying only on the regularity of the linguistic environment and basic memory and attention processes, this may be an important method of learning nouns for infants, and even adult travelers in a foreign country.

In adult cross-situational learning studies (e.g., Yu & Smith 2007), participants are asked to learn the meaning of alien words by watching a series of training trials. On each trial learners see an array of unfamiliar objects (e.g., four sculptures) and hear pseudowords (e.g., *stigson, bosa*). The meaning of each pseudoword is ambiguous on a given trial, because although each word refers to a single onscreen object, the intended referent is not indicated. In a typical learning scenario, participants attempt to learn 18 word-object pairings from 27 trials, with four words and four objects given per trial. In this design, each word-referent pair is presented six times over the five-minute training period. Learning a correct word-object pairing requires some form of accumulation of word-object co-occurrences. When tested on each word and given four trained objects to choose from, participants chose the correct object for half of the 18 words, on average (Yu & Smith, 2007).

However, learning environments in the real world are likely not as simple: there may be objects present that go unnamed, some spoken words (e.g. articles) do not refer to particular objects, and object names may be spoken when the intended object is not visible. These forms of noise likely interfere with learning to some extent. When a word is heard without the object it previously co-occurred with several times, is a learner to map it to a new object? What if that object already has a name? Conversely, when an object is seen, but the word it previously occurred with is not heard, will learners lose certainty about the old mapping, and even associate a new word with it?

In this study, we take baseline conditions from Yu & Smith (2007) that present an equal number of words and objects on each trial and either add or remove words or objects in a systematical way in order to change various co-occurring statistics that learning may rely on. We investigate several critical factors that matter to learning, such as conditional probability of words given objects during learning, final test probability, and contextual diversity—the number of other pairs each pair appears with (Kachergis, Yu, & Shiffrin, 2009b). Following Fazly, Alishahi, and Stevenson (2010), we also investigate additional two factors – age of exposure (i.e., when a pair first appears) and context familiarity (the mean frequency of the objects a given pair appears with). Not only are these factors likely to influence how well people learn, but likely so will the fact that the trials contain an unequal number of words and objects. Previous studies have also typically presented an equal number of words and objects on each trial, which may induce participants to only consider 1-to-1 mappings (although see Vouloumanos, 2008 as well as mutual exclusivity investigations: Kachergis, 2012; Ichinco, Frank, & Saxe, 2009; Yurovsky & Yu, 2008).

Finally, we use a recent associative model of cross-situational learning (Kachergis, Yu, & Shiffrin, 2012) to shed light on differences between the conditions. The model assumes that learners have access to both their familiarity and their uncertainty about the word-object pairings present on a given trial, and that attention competes for uncertain stimuli and for already-strong pairings. This model matches adult behavior in a number of previous cross-situational experiments (Kachergis, 2012; Kachergis, Yu, & Shiffrin, 2013).

Experiment

Participants were asked to learn 18 word-referent pairs from a series of individually ambiguous training trials using the cross-situational word learning paradigm (Yu & Smith, 2007). Each training trial was comprised of a display of two or more novel objects and two or more spoken pseudowords, depending on condition. With no indication of which word refers to which object, on a single trial, learners can only guess at the correct word-referent mappings. However, since words always appear on trials with their intended referents, the correct pairings may be learned over the series of trials.

In this study, many conditions were created by manipulating training conditions from Yu and Smith (2007)—the 2x2 (i.e., 2 word-object pairs per trial), 3x3, and 4x4 conditions— to introduce different types of noise which is arguably more in line with real-world learning, such as a non-referential word, an unnamed object, or both. In every condition, participants experienced a series of training trials with a total of 18 intended word-object pairs. The same pair was never allowed to appear in neighboring trials in conditions, as previous studies have shown such temporal contiguity improves learning (Kachergis, Yu, & Shiffrin, 2009a; Kachergis, Yu, & Shiffrin, 2013). In the baseline **2x2** (54 trials), **3x3** (36 trials), and **4x4** (27 trials) conditions, each word and object appear 6 times. Every time a given object is present, the intended word is presented (p(w|o)=1), and every time a given word is presented, the intended object is present (p(o|w)=1). Most conditions in Table 1 were built from these three baseline conditions. We manipulate the number of words and objects per trial, thus changing their frequency. This also changes the probability of hearing the word for a given object on a trial (in Table 1, Trial p(w|o)). The probability of seeing an object given that its label was heard was always 1 (Trial p(o|w)). Test p(o|w) in Table 1 shows the probability of guessing the intended object for a given word after experiencing all of the training.

In the **2x4** condition, words appeared 6 times and objects 12 times, so on each trial the probability of hearing the intended word for a given object is p(w|o)=.5. In the **2x3** condition, objects appear 9 times, making p(w|o)=.67. In the **2x4** condition, each word appears 6 times and each object appears 12 times. In the **3x3 +1w/o** condition, an additional random word and object were shown on each trial. In the **4x4 +2w/o** condition, two additional random words and objects were shown per trial. In the **3x4** condition, each word appears 6 times, each object 8 times (p(w|o)=.75). In the **3x4 1/.5** condition, words appear 6 times, and 12 objects appear only with their words (p(o,w)=1), while 6 objects appear 12 times (p(w|o)=.5). In the **3x4 1/.66** condition, words appear with their objects 6 times (p(w|o)=1), but 12 objects appear 3 additional times (p(w|o)=.66) without their words. In the **3x4 +6o** condition, 18 word-object pairs co-occur 6 times, and 6 additional objects occur as noise.

The **1x3** condition divided each trial of the 3x3 condition into 3 trials with one word and 3 objects, and shuffled the trials so no objects (or words) repeated trial-to-trial. Thus, words appeared 6 times, and objects 24 times (p(w|o) = .33). The **1x4** condition divided the 4x4 trials as 1x3 did for 3x3, meaning that objects appeared 24 times (p(w|o) = .25).

Calculated for each item per condition, Table 1 also shows the average "Age" of Exposure (trial the pair first appears), Context Familiarity (defined by Fazly, Alishahi, and Stevenson (2010) as the mean co-occurrence with all other pairs across exposures), and Context Diversity (the number of unique pairs a pair co-occurs with over training).

Condition	Word Freq.	Object Freq.	Trial p(w|o)	Test p(o|w)	Context Familiarity	"Age" of Exposure	Context Diversity	Num. Subjs.	Correct
2x2	6	6	1	0.5	3.5	5.6	5.1	19	0.79
2x3	6	9	0.67	0.33	3.3	6.0	9.1	55	0.56
2x4	6	12	0.5	0.25	3.3	5.6	11.8	33	0.30
3x3 +1w/o	9	9	1	0.22	5.1	4.3	12.9	39	0.17
4x4 +2w/o	12	12	1	0.12	6.6	3.8	16.2	39	0.10
3x3	6	6	1	0.33	3.5	3.7	8.8	36	0.43
1x3	6	18	0.33	0.33	3.2	17.5	8.7	63	0.52
3x4	6	8	0.75	0.25	3.4	3.7	12.3	25	0.19
3x4 +6o	6	6	1	0.25	3.5	3.7	13.6	20	0.27
3x4 1/.5	6	8	1 / .5	0.25	4.3	3.7	11.3	25	0.22
3x4 1/.66	6	8	1 / .66	0.25	3.6	3.7	12.1	25	0.21
4x4	6	6	1	0.25	3.5	2.8	12.2	77	0.31
1x4	6	24	0.25	0.25	3.1	19.9	12.0	40	0.19

Table 1. Summary of conditions in the Experiment.

Subjects

Participants were undergraduates at Indiana University who received course credit for participating. The number of participants in each condition are shown in Table 1 (Num. Subjs. column). None had participated in previous cross-situational experiments.

Stimuli

Each training trial consisted of an array of 2-4 uncommon objects (e.g., sculptures) and 2-4 spoken pseudowords, depending on condition (see Table 1). The computer-generated pseudowords are phonotactically-probable in English (e.g., "bosa"), and were spoken by a monotone, synthetic female voice. The words and objects used for each condition were drawn from a set of 72 words and 72 objects.

Training for each condition consisted of between 27 and 108 trials. Each training trial began with the appearance of two to four objects (differing by condition) which remained visible for the entire trial. After 2s of initial silence, each word was heard (1s per word, with 2s of silence after each).

Procedure

Participants were told they would see a series of trials with some objects and alien words, but that the words would be presented in random order. They were also told that their knowledge of which words belong with which objects would be tested at the end.

After each training block, participants' knowledge of word-object mappings was assessed using 18-alternative forced choice (18AFC) testing: on each test trial a single word was played, and the participant was instructed to choose the appropriate object from a display of all 18 trained objects. Each of the 18 words was tested once in a random order.

Results & Discussion

As shown in Fig. 1, all of the conditions had mean performance significantly above chance (18AFC chance = .056). The 2x2 baseline condition had by far the highest performance (*M*=.79). Adding another object to each trial—without it's intended word—harmed learning (2x3: *M*=.56). 2x4 adds yet another object, further decreasing both Trial *p*(w|o) and Test *p*(o|w), resulting in even lower performance (*M*=.30). Adding an extra pair (3x3 +1w/o) or two (3x3 +2w/o) is even more harmful (*M*=.17, *M*=.10, resp.); it both lowers Test *p*(o|w) and creates more possible pairings to consider on each trial. For another example, the 1x3 and 1x4 conditions are identical in all of the other factors except that there were 1 word and 3 objects in the 1x3 condition (0.33) but 1 word and 4 objects in the 1x4 condition (0.25). This one change caused a dramatic performance difference from *M*=.53 to *M*=.19. Meanwhile, it may not seem like there is a dramatic difference between the 1x3 and 2x3 conditions. All this suggest that given multiple factors that can be used to characterize statistical information in training data, and the flexibly of human statistical learning systems, it is difficult to pull apart all of the effects in terms of conditions—especially in the 3-word and 4-word conditions—as a change in one factor is often correlated with changes in several other factors (e.g., contextual diversity and context familiarity).

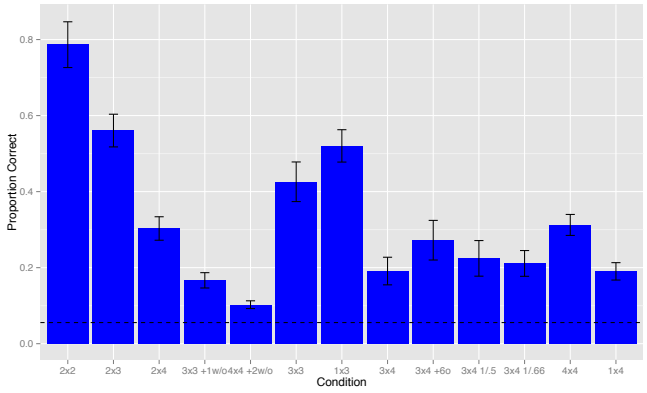

Figure 1: Mean accuracy by training condition. Performance was variable, but all conditions were above chance (18AFC=.056). Error bars show +/-SE.

To better understand the effects of the various factors, we fit a logistic mixed-effects regression model to the trial-level accuracy data using the lme4 package in R (Bates and Maechler, 2010; R Development Core Team, 2010). Mixed logit models are more appropriate for forced-choice data than ANOVAs, especially when different conditions yield different amounts of data, as in the present experiment (Jaeger, 2008). The model included subject as a random factor, and Trials/Condition, Word Frequency, Object Frequency, Trial *p*(w|o), Test *p*(o|w), Contextual Diversity, Age of Exposure, and Context Familiarity as fixed factors. All of these factors were scaled to remove collinearity. Shown in Table 2, there was a significant negative intercept, showing that participants were less likely to choose the correct answer than the incorrect answer. Trials/Condition and Test *p*(o|w) both had significant, large positive coefficients (.75 and .78), showing that longer training periods were better, as well as stronger correct associations—both of which occur more in the conditions with fewer pairs per trial (i.e., 2x2 rather than 4x4).

Factor	Coefficient	*Z*	p-value
(Intercept)	-0.75	-9.20	<.001
Trials/Cond	0.75	4.57	<.001
Word Freq	-0.10	-0.92	=.36
Obj Freq	-0.58	-2.75	<.01
Trial *p*(w|o)	-0.14	-0.88	=.38
Test *p*(o|w)	0.78	5.67	<.001
Cont. Fam.	0.20	2.82	<.01
Age of Exp	-0.08	-1.93	=.05
Cont. Div.	0.17	2.24	<.05

Table 2. Summary of logistic regression results.

Word frequency did not contribute significantly to correctness, but object frequency had a negative coefficient, showing that additional repetitions of an object on trials without the intended word indeed inhibited learning of that object. Trial $p(w|o)$ was not a significant predictor of accuracy; it seems the other (partially-correlated) factors better capture the variance. Context Familiarity and Contextual Diversity both have significant positive coefficients (.20 and .17). Though they are correlated (r=.56), these two factors both help people learn words. Age of Exposure had a marginally significant negative coefficient (-.08), showing that earlier-appearing pairs are indeed a bit more likely to be learned.

In total, these results offer a look at the factors that influence cross-situational word learning, and an estimate of their relative magnitudes. We now apply a recent associative model of cross-situational word learning to see whether it can account for word-learning in these noisy environments, and to see whether the recovered parameters yield any additional insight.

Model

To better understand how the condition demands differ, we introduce an associative model of cross-situational word learning proposed by Kachergis, Yu, and Shiffrin (2012a).

The model assumes that learners do not equally attend to all word-object pairings on a trial (i.e., store all co-occurrences). Rather, selective attention on a trial is drawn to strengthen associations between words and objects that have co-occurred previously. This bias for familiar pairings competes with a bias to attend to stimuli that have no strong associates (e.g., as a novel stimulus). The competing familiarity and uncertainty biases allow the model to exhibit fast mapping, since a novel word-novel object combination will demand more attention, and mutual exclusivity: a novel word will only become weakly associated with an already-known referent (Kachergis, Yu, & Shiffrin, 2012). For example, suppose word w_1 and object o_1 have appeared together and are thus somewhat associated, while w_7 and o_7 are novel. Given a trial with both pairs: $\{w_1,o_1,w_7,o_7\}$, w_1-o_1 demands more attention than w_7-o_1, w_1-o_7, or w_7-o_7, since w_1-o_1 is stronger than baseline. However, attention is also pulled individually to w_7 and to o_7, since both of these novel stimuli have no strong associates. Uncertainty is measured by the entropy of each stimulus' association strengths. Because of the high joint uncertainty of w_7 and o_7, more attention is given to the association w_7-o_7. Thus, attention is mostly divided between w_1-o_1 and w_7-o_7, although the other pairings will be strengthened a bit.

Formally, let M be an n word × n object association matrix that is incrementally built during training. Cell $M_{w,o}$ will be the strength of association between word w and object o. Strengths are subject to forgetting (i.e., general decay) but are augmented by viewing the particular stimuli. Before the first trial, M is empty. On each training trial t, a subset S of m word-object pairings appears. If new words and objects are seen, new rows and columns are first added. The initial values for these new rows and columns are k, a small constant (here, 0.01).

Trial-to-trial, association strengths decay and then a fixed amount of associative weight, χ, is distributed among the presented word-object associations and added to the strengths. The rule used to distribute χ (i.e., attention) balances a bias for attending to unknown stimuli with a bias for strengthening already-strong associations. When a word and referent are repeated, extra attention (i.e., χ) is given to this pair: a prior knowledge bias. Stimuli with no strong associates also attract attention, whereas pairings between uncertain objects and known words, or vice-versa, draw little attention. Stimulus uncertainty is measured by entropy (H), which is 0 when the outcome of a variable is certain (e.g., a word appears with one object, and has never appeared with any other object), and maximal ($log_2 n$) when all of the n possible object (or word) associations are equally likely (e.g., when a stimulus has not been observed before, or if a stimulus were to appear with every other stimulus equally). In the model, on each trial the entropy of each word (and object) is calculated from the normalized row (column) vector of associations for that word (object), $p(M_{w,\cdot})$, like so:

$$H(M_{w,\cdot}) = -\sum_{i=1}^{n} p(M_{w,i}) \cdot \log(p(M_{w,i}))$$

The update rule for allocating attention and adjusting strengths for the stimuli presented on a trial is:

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

In this equation, α is a parameter governing forgetting, χ is the weight being distributed, and λ is a scaling parameter governing differential weighting of uncertainty and prior knowledge (familiarity). As λ increases, the weight of uncertainty (i.e., the exponentiated entropy term, which includes both the word's and object's association entropies) increases relative to familiarity. The denominator normalizes the numerator so that exactly χ associative weight is distributed among the potential associations on the trial. Only decay operates for stimuli not on the current trial. After training, a learner is tested with each word and chooses an object from n alternatives in proportion to the association strengths of each alternative to that word.

In sum, this associative model learns trial-by-trial by distributing attention in a way that corresponds with both our intuitions about word-learning—using competing biases for familiar pairings and uncertain stimuli—and a number of empirical findings. Three parameters (χ, α, and λ) were fit using log likelihood to each individual's choices in each training condition. The model achieved quite a good fit to the data, with R^2=.98. Figure 2 shows mean model performance for individuals' model fits by condition. Figure 3 shows individuals' mean performance in each condition versus the model's performance. Next, we investigate the

parameter values for each condition to see what they tell us about the cognitive effects of different types of noise.

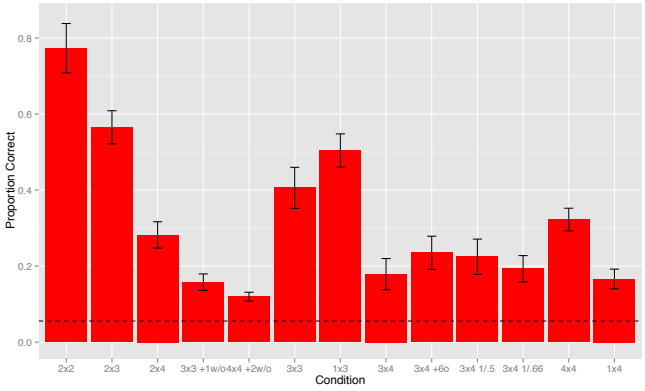

Figure 2. Model performance closely matches human performance (Fig. 1) and variability in the Experiment.

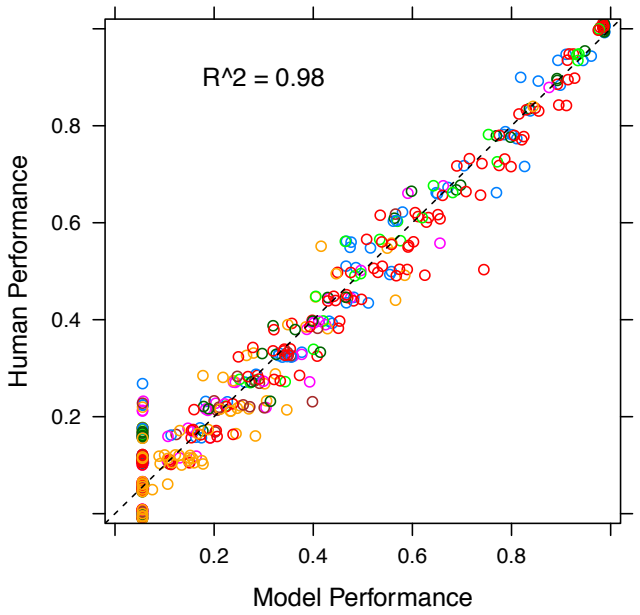

Figure 3. Individual performance versus model fit: the model was capable of closely matching the behavior of most participants.

We first looked at correlations between each parameter and performance. χ was positively correlated with learning (Pearson's r=.72, t(494)=22.74, p<.001), which is consistent with our interpretation of χ as a learning rate; how much associative weight can be distributed per trial. λ was negatively correlated with performance (r = -.22, t(494)=-5.04, p<.001): greater focus on uncertain stimuli seems to harm learning, at least in the conditions of this Experiment. λ and χ were also negatively correlated (r = -.20, t(294)=-4.64, p<.001), meaning that uncertainty-focused learners tended to have slower learning rates. All other correlations were <|.03|, and not significant.

We also investigated whether there were differences in parameters by condition. Ideally, the parameters of a cognitive model should be cognitively interpretable. For example, in our model, χ is for now a learning rate per trial, but should likely depend on how many possible associations there are on a trial and how much time there is to consider them. If systematic differences in particular parameters were required to fit performance in some of the conditions, then we may be able to pinpoint which factors learning rate and memory decay depend on, and redefine them in more meaningful units. An ANOVA by condition for each parameter showed significant differences for all three parameters (χ: F(12,482)=11.63, p<.001; λ: F(12,482)=2.13, p=.01; α: F(12,482)=2.70, p<.01). Table 3 shows the mean parameters found for each condition. We emboldened the highest mean values for each parameter and italicized the lowest in order to highlight the conditions with unusual mean parameter values.

For χ, the 2x2 condition has the highest value (19.47), and this condition also yields the highest performance in humans. 2x2 also has the lowest λ (i.e., more focus on familiarity) and α (i.e., faster decay), the latter of which may mitigate the high learning rate a bit. Conditions with the next-highest learning rates—2x3 (9.87) and 1x3 (10.32)—had the next-highest performance (.56 and .52). 1x3, along with 1x4 also had the highest mean α = .94 (memory fidelity). These two conditions have the shortest trials (5s), along with the fewest possible associations: only one word and three or four objects. The conditions with the lowest learning rates, 3x4 (χ=.35) and 1x4 (χ=.47), have fairly low performance (.19 and .19). The short trial time for the 1x4 condition may not give subjects enough time to pick out the single correct pairing.

Condition	Correct	χ	λ	α
2x2	0.79	**19.47**	*5.0*	*0.85*
2x3	0.56	9.87	6.9	0.92
2x4	0.30	1.73	9.3	0.88
3x3 +1w/o	0.17	0.91	8.6	0.89
4x4 +2w/o	0.10	3.01	8.7	0.87
3x3	0.43	6.30	6.1	0.90
1x3	0.52	10.32	7.3	**0.94**
3x4	0.19	*0.35*	7.5	0.89
3x4 +6o	0.27	5.07	**9.2**	0.89
3x4 1/.5	0.22	0.99	8.0	0.92
3x4 1/.66	0.21	1.58	**9.1**	0.88
4x4	0.31	2.80	7.9	0.87
1x4	0.19	*0.47*	**9.1**	**0.94**

Table 3. Mean of best-fitting parameters for each condition. The largest and smallest mean values of each parameter are emboldened and italicized, respectively.

In the 3x4 condition, there are again more objects than words, and many possible associations. The other 3x4 conditions also had low performance and low learning rates,

except for 3x4 +6o, in which participants may have had little difficulty ignoring the extraneous objects (which are less confusing since they are occur infrequently, and never with a consistent name). It is hard to see a pattern in λ, the relative focus on uncertainty vs. familiar pairings (roughly, explore vs. exploit). We do not yet have any reason to believe λ should remain fixed; learners may well change it—implicitly or strategically—depending on task demands. Moreover, previous investigations found that λ has little effect on the shape of learning curves (Kachergis, Yu, & Shiffrin, 2012b).

Discussion

In the language environment, many objects in a scene may go unlabeled, whether they are novel or familiar. For the sake of simplicity, previous studies of cross-situational learning presented an equal number of words and objects on each trial, and a word's intended referent was always present (and vice-versa; e.g. Yu & Smith, 2007; Kachergis, Yu, & Shiffrin, 2009a, 2009b). In this study, we presented learners with a variety of conditions with different kinds and degrees of statistical noise (e.g., extra objects, mismatched words and objects). Although performance varied widely in different conditions, learners performed significantly above chance in all conditions.

To better understand what factors influence learning, we measured various statistics about items in each condition, and tried to predict learning from these statistics. Greater contextual diversity—how many pairs a pair appears with during training, context familiarity—the average frequency of pairs a pair appears with, trials per condition, and overall strength of the correct pairing all significantly improved the odds of learning a pair. Being exposed to a pair earlier in training improved learning of that pair, but being exposed to an object more often inhibited learning, because in this study extra occurrences of an object were likely to be noise (e.g., appearing on a trial where it goes unnamed). These conditions and measures provide important constraints for word-learning models, as well as demonstrating that cross-situational learning is robust under a variety of types of noise.

We applied a recent associative word-learning model to these data, and found that it could account for individuals' behavior in each of the conditions. We investigated the average parameter values for individuals in each condition, and found that they differed. The learning rate parameter was strongly linked to overall performance, and was high when there were few pairings to consider on each trial (e.g., 2x2, 1x3)—unless most of them were noise, and presented quickly (e.g., 1x4). There less memory decay in conditions with very one word per trial, and thus few associations (1x3, 1x4), although the most decay occurred in the 2x2 condition, but that was balanced by the fast learning rate. Overall, we have a somewhat clearer idea of what the model's parameters do under different noise conditions, but we do not yet have a wholly satisfactory psychological interpretation of them.

In summary, cross-situational learning is robust under a many noise conditions that more closely resemble situations learners may encounter in the real world than in previous studies. Moreover, we have presented a large dataset that we hope will inspire new experiments to test the limits of cross-situational learning, and will constrain and inform modeling efforts.

References

Bates, D. & Maechler, M. (2010). lme4: Linear mixed-effects models using S4 classes. R package version 0.999375-37. http://CRAN.R-project.org/package=lme4

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

Gleitman, L. (1990). The structural sources of verb meanings. *Language Acquisition*, *1*(1), 1–55.

Ichinco, D., Frank, M.C., & Saxe, R. (2009). Cross-situational word learning respects mutual exclusivity. In N. Taatgen, H. van Rijn, J. Nerbonne, & L. Schomaker (Eds.) *Proceedings of CogSci 31* (pp. 749–754).

Jaeger, T. F. (2008) Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*, 434-446.

Kachergis, G., Yu, C., & Shiffrin, R. M. (2009a). Temporal contiguity in cross-situational statistical learning. In N. Taatgen, H. van Rijn, J. Nerbonne, & L. Schomaker (Eds.) *Proceedings of CogSci 31*. Austin, TX: Cognitive Science Society.

Kachergis, G., Yu, C., & Shiffrin, R. M. (2009b). Frequency and contextual diversity effects in cross-situational word learning. In N. Taatgen, H. van Rijn, J. Nerbonne, & L. Schomaker (Eds.) *Proceedings of CogSci 31* (pp. 755-760).

Kachergis, G., Yu, C., & Shiffrin, R.M. (2012a). An associative model of adaptive inference for learning word-referent mappings. *Psychonomic Bulletin & Review*, *19*(2), 317-324.

Kachergis, G., Yu, C., & Shiffrin, R. M. (2012b). Cross-situational word learning is better modeled by associations than hypotheses. IEEE Conference on Development and Learning / EpiRob 2012.

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

Markman, E.M. & Wachtel, G.F. (1988). Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive Psychology*, *20*, 121–157.

Vouloumanos, A. (2008). Fine-grained sensitivity to statistical information in adult word learning. *Cognition*, *107*(2), 729-742.

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

Yurovsky, D. & Yu, C. (2008). Mutual exclusivity in cross-situational statistical learning. *Proceedings of CogSci 30* (pp. 715–720). Austin, TX: Cognitive Science Society.