Cross-Situational Learning: An Experimental Study of Word-Learning Mechanisms

Kenny Smith,[a] Andrew D. M. Smith,[b] Richard A. Blythe[c]

[a]*Language Evolution and Computation Research Unit, School of Philosophy, Psychology and Language Sciences, University of Edinburgh*
[b]*Department of English Studies, University of Stirling*
[c]*SUPA, School of Physics and Astronomy, University of Edinburgh*

Abstract

Cross-situational learning is a mechanism for learning the meaning of words across multiple exposures, despite exposure-by-exposure uncertainty as to the word's true meaning. We present experimental evidence showing that humans learn words effectively using cross-situational learning, even at high levels of referential uncertainty. Both overall success rates and the time taken to learn words are affected by the degree of referential uncertainty, with greater referential uncertainty leading to less reliable, slower learning. Words are also learned less successfully and more slowly if they are presented interleaved with occurrences of other words, although this effect is relatively weak. We present additional analyses of participants' trial-by-trial behavior showing that participants make use of various cross-situational learning strategies, depending on the difficulty of the word-learning task. When referential uncertainty is low, participants generally apply a rigorous eliminative approach to cross-situational learning. When referential uncertainty is high, or exposures to different words are interleaved, participants apply a frequentist approximation to this eliminative approach. We further suggest that these two ways of exploiting cross-situational information reside on a continuum of learning strategies, underpinned by a single simple associative learning mechanism.

Keywords: Word learning; Cross-situational learning; Associative learning

1. Introduction

Determining the meaning of a newly encountered word should be extremely hard, due to the (in principle, unlimited) referential uncertainty inherent in the task (Quine, 1960).

Correspondence should be sent to Kenny Smith, Language Evolution and Computation Research Unit, School of Philosophy, Psychology and Language Sciences, University of Edinburgh, Dugald Stewart Building, 3 Charles Street, Edinburgh, EH8 9AD, UK. E-mail: kenny@ling.ed.ac.uk

Despite this, children are prodigious and rapid word learners, learning around 60,000 words by age 18 (Bloom, 2000), and capable of identifying at least some aspects of the meaning of a novel word after only a few exposures, through so-called *fast mapping* (Carey & Bartlett, 1978; see Horst & Samuelson, 2008 for review). Much word learning research has focused on how referential uncertainty can be reduced, by eliminating from consideration meanings which are theoretically possible, but in practice spurious. Socio-pragmatic, representational, interpretational, and syntactic heuristics have been proposed: for example, children use behavioral cues to identify the speaker's attentional focus (Baldwin, 1991; Nappa, Wessel, McEldoon, Gleitman, & Trueswell, 2009; Tomasello & Farrar, 1986); they assume words refer to whole objects, rather than their parts or properties (Macnamara, 1972); they exploit their knowledge of other word meanings, for example, by assuming that words have mutually exclusive meanings (Markman & Wachtel, 1988); argument structure and syntactic context facilitate word learning (Gillette, Gleitman, Gleitman, & Lederer, 1999; Gleitman, Cassidy, Nappa, Papafragou, & Trueswell, 2005). These heuristics all act to restrict referential uncertainty, but they are unlikely to eliminate all ambiguity on every word learning exposure: Some residual uncertainty will remain.

Cross-situational learning (henceforth XSL, e.g., Pinker, 1989, 1994; Gleitman, 1990) is a mechanism for word learning despite referential uncertainty. In each exposure to a word, the context (both linguistic and nonlinguistic) in which the word is used, together with the learner's word-learning heuristics (of the sort outlined above), provides a set of multiple candidate referents. Although this means that the referent of a word cannot be identified on a single exposure, a learner who can combine information across multiple exposures can determine the most probable referent, by intersecting the various sets of candidate referents. As it does not require the elimination of all uncertainty, XSL allows words to be learned by a learner lacking the sophisticated and (presumably) cognitively demanding inferential processes needed to eliminate referential uncertainty entirely.

Computational models suggest that XSL can be used to accurately infer the meanings of words from small but realistic corpora of language use (Frank, Goodman, & Tenenbaum, 2009; Siskind, 1996; Yu, 2008; Yu, Ballard, & Aslin, 2005), and we have shown mathematically that XSL can allow large, language-scale lexicons to be learned in the face of considerable referential uncertainty (Blythe, Smith, & Smith, 2010). A growing body of experimental evidence also suggests that cross-situational learning of small numbers of words despite exposure-by-exposure referential uncertainty may be within the capabilities of both adults (Gillette et al., 1999; Smith, Smith, & Blythe, 2009; Xu & Tenenbaum, 2007b; Yu & Smith, 2007) and children (Akhtar & Montague, 1999; Childers & Pak, 2009; Piccin & Waxman, 2007; Smith & Yu, 2008; Xu & Tenenbaum, 2007a,b). However, experiments where referential uncertainty is high enough to render XSL impossible are rare (but see K. Smith et al., 2009). Understanding the limits of XSL with respect to referential uncertainty is important, as it provides an indirect way to identify the strength of the word-learning heuristics discussed above: In order for word learning to be possible at all, these heuristics must reduce referential uncertainty to levels which render XSL possible. Here, we describe an experiment designed to extend our understanding of the mechanisms and limits of XSL. We investigate how well humans can learn word meanings using

cross-situational information under different levels of referential uncertainty and different modes of presentation, and we provide a novel technique for the detailed exploration of how learners exploit cross-situational information.

We present an experiment showing that word learning deteriorates with increasing referential uncertainty, but it is still possible at levels of referential uncertainty more than double those previously tested. We find some evidence that words presented through exposures interleaved with exposures to other words are harder to learn than those presented consecutively, both in terms of learning success and in terms of learning time. We also analyze how participants' learning behavior changes according to task difficulty. Although humans are indeed effective cross-situational learners, even under relatively high referential uncertainty, the rigor with which they exploit cross-situational information is modulated by the degree of referential uncertainty and presentation mode: Full eliminative XSL is only used under low levels of referential uncertainty and consecutive presentation; participants shift to a frequentist approximation as the task becomes harder (as referential uncertainty increases, or when exposures are interleaved). We then discuss the strengths and weaknesses of our approach and sketch how the various flavors of XSL can be accounted for by a single underlying associative learning mechanism.

2. Experiment

We adopt a paradigm combining the repeated testing approach of Gillette et al. (1999) with the controlled and quantified level of referential uncertainty of Yu and Smith (2007) and Smith and Yu (2008): Participants are repeatedly exposed to a small set of word-object pairings, with each training exposure immediately followed by a test requiring participants to identify which referent object they think the word refers to. We explore the impact on learning of both the degree of referential uncertainty and the interleaving of exposures to multiple words.

2.1. Method

2.1.1. Participants

We recruited 48 participants (34 females) aged 18–42 ($M = 23.55$) through the University of Edinburgh Careers Service database. Each participant was paid £5 for his or her participation.

2.1.2. Materials

We produced 120 novel objects, consisting of a mix of photographs of unusual real-world objects (e.g., a bicycle light retaining clamp) and artificial objects created by cutting and pasting together component parts of pictures of technological artefacts. Two lists of eight nonsense words were created (using the English Lexicon Project Website: Balota et al., 2007): The words followed English phonotactics and were all stressed on the first syllable, but they varied according to the number of syllables (1, 2, or 3) and word onset (vowel,

single consonant, or consonant cluster).[1] Spoken forms were produced using the Victoria voice on the Apple Mac OS X speech synthesizer. The experiment was developed using Slide Generator (http://www.psy.plymouth.ac.uk/research/mtucker/slidegenerator.htm), and participants were tested at computers running Windows XP, providing responses via a mouse.

2.1.3. Procedure

Participants were asked to learn the names of eight novel objects. They were briefed that each object would be named repeatedly, and that several objects might be present on each presentation; they were not explicitly instructed to apply a cross-situational approach. Target word forms were selected at random without replacement from one of the word lists; each target word had an associated set of 15 referent objects, selected at random and without replacement from the larger set of 120 novel objects. The target referent and nontarget context items for a given word were selected from this set of 15 objects: As there was no overlap between the sets of referent objects for different words, participants could not use mutual exclusivity (Markman & Wachtel, 1988) or similar heuristics to reduce the referential uncertainty of subsequently presented words. The first two words encountered by each participant were designated practice words, to familiarize participants with the task and the experimental interface, and were ignored in the analysis.[2] The remaining six words were organized in two blocks of three words; they varied in referential uncertainty (quantified in terms of the *context size*, C, namely $C = 2$, 5, or 8 nontarget referents co-present with the target referent on each exposure, see also Table 1) and mode of presentation (in the *consecutive* block, all exposures to a word were presented consecutively; in the *interleaved* block, exposures to one word were interleaved with exposures to the other two words, in strict rotation). Each participant experienced each level of C twice, once in consecutive presentation and once interleaved.[3]

Each exposure to a word consisted of two parts (see Fig. 1 for example):

1. *Training*, in which participants heard the word being spoken through headphones, while several (the target + C) objects were simultaneously presented on screen. The training screen was presented for 5 s.

Table 1
Parameters: Mode of presentation, number of distractors (C), learning time for a perfect cross-situational learner to learn the word (e'), and the number of training-test exposures presented (e^{max})

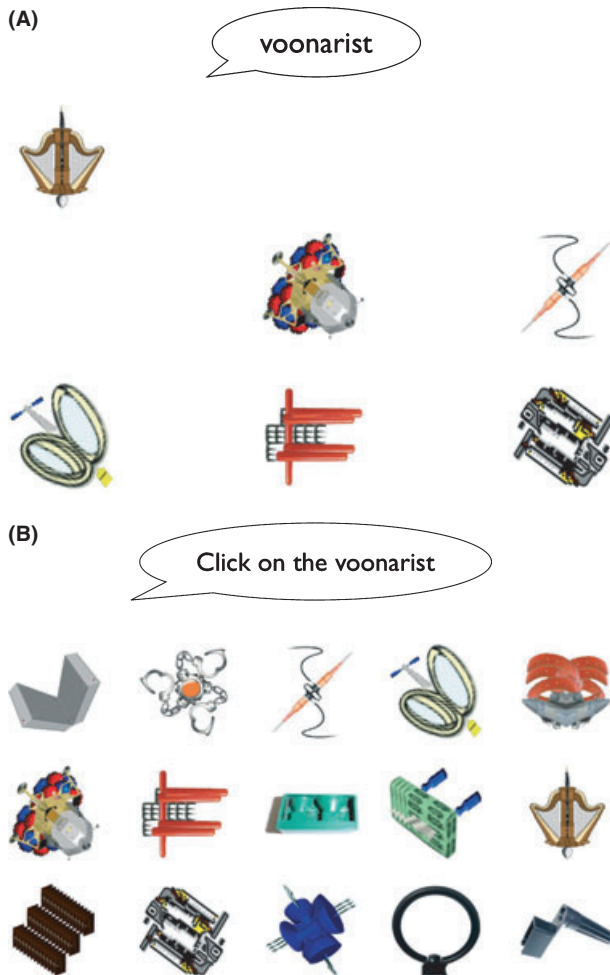|	Presentation	C	e'	e^{max}
Practice Word 1	Consecutive	1	2	6
Practice Word 2	Consecutive	2	2	6
Word 1	Consecutive	2	4	12
Word 2	Consecutive	5	4	12
Word 3	Consecutive	8	4	12
Word 4	Interleaved	2	4	12
Word 5	Interleaved	5	4	12
Word 6	Interleaved	8	4	12

Fig. 1. A single train-test exposure. (A) Training. Participants are presented visually with the target and several (here: five) nontarget referents, paired with an aural presentation of a nonsense word (here: *voonarist*). (B) Testing. Participants are immediately aurally prompted to select the referent corresponding to the nonsense word, from an array of 15 possibilities.

2. *Testing*, immediately following the training screen, where participants were presented with an array of 15 objects, and asked to click on the one they thought the word referred to. Participants had a maximum of 30 s to respond.[4] The test array contained all 15 possible referents for a word, with position in the test array being constant across exposures to a given word.

Practice words 1 and 2 were always presented first. The choice of word list (list 1 or list 2), order of presentation of the two blocks (consecutive or interleaved first), and the order in which the three levels of referential uncertainty were encountered within a block (six

possible orderings)[5] were counterbalanced across participants. To reduce between-subjects manipulations, the same presentation order was used across blocks for a given participant (e.g., if they received the ordering $C = 2$, $C = 5$, $C = 8$ in their first block, this ordering was repeated in their second block), yielding 24 combinations (2 word lists × 2 block orders × 6 orders of levels of C); two participants were run for each such condition.

We designed the training sequences with participants organized into yoked pairs. Within a yoked pair, for a given value of C, identical training data and test arrays were used, but the participants in the pair differed in whether they received those exposures via consecutive or interleaved presentation. This allows an additional by-pairs analysis on the effects of interleaving on learning. The sequence of exposures for a particular word and yoked pair was generated at random by a custom-written program, so that a perfect eliminative cross-situational learner would learn the word after $e' = 4$ exposures.[6]

2.2. Results

A word is defined as *learned* if the target referent was chosen on the final test exposure. A word is learned on exposure e if the target referent was chosen on exposure e and on all subsequent exposures; the learning time for a word is the smallest such e. In the following sections we present results for learning success, learning times for successful learners, and the strategies employed.

2.2.1. Learning success

Table 2 shows the number of participants who successfully learned each word, together with the number of words we would expect to have been learned if learners were using the best possible non-XSL strategy, achieved by simply choosing randomly from all referents in the current context (i.e., in the training exposure immediately before the test). With this strategy, which we call *Random from $\mathscr{C}$*, a learner would learn a given word with probability $\frac{1}{(C+1)}$. Comparing the observed success rates with Random from $\mathscr{C}$ provides a direct test for XSL: If this baseline is exceeded, then XSL must be taking place. Testing against a weaker baseline (e.g., random selection from the test array, as in Yu & Smith, 2007) does not conclusively demonstrate XSL, as discussed in K. Smith et al. (2009). In all cases, the observed

Table 2

Number of participants learning words in each experimental condition (out of 48), compared with the best possible noncross-situational learning strategy (Random from $\mathscr{C}$)

	C		
Presentation	2	5	8
Consecutive	46***	37***	31***
Interleaved	44***	37***	23***
Random from $\mathscr{C}$	16	8	5.33

Note. Learning success that is significantly greater than Random from $\mathscr{C}$ is indicated by asterisks (***, $p < .001$).

learning success rates[7] are significantly higher than the Random from $\mathscr{C}$ baseline (smallest $\chi^2(1) = 65.84$, $p < .001$, occurring in the interleaved, $C = 8$ condition), thus demonstrating that participants are integrating information cross-situationally.

To evaluate the effect of referential uncertainty and presentation mode on learning success, we fit a Cox proportional-hazards regression model (Cox, 1972):[8] This very general regression model allows us to model interindividual differences in learning success and does not rely on any assumptions concerning the shape of the underlying distribution of event times, but instead assumes that the underlying hazard rate is a function of the independent covariates (the within-participant predictor variables). The Cox model provides an estimated hazard ratio (HR) indicating the relative likelihood of word learning in an experimental group compared with a control (Spruance, Reid, Grace, & Samore, 2004).

This regression analysis shows a significant effect for C after adjustment for subject effects (relative to the $C = 2$ baseline: $C = 5$, HR = 0.374, $p < .001$; $C = 8$, HR = 0.192, $p < .001$). These hazard ratios indicate that words in the $C = 5$ condition are at approximately one third the baseline ($C = 2$) ''risk'' of being learned at any given exposure, and $C = 8$ words are at approximately one fifth the baseline ''risk'' of being learned. The difference between $C = 5$ and $C = 8$ is also significant (relative to a $C = 5$ baseline: $C = 8$, HR = 0.514, $p < .001$). The model also indicates a significant effect for presentation mode (relative to the consecutive presentation baseline: interleaved, HR = 0.72, $p = .032$), and no interaction between referential uncertainty and presentation mode ($p \geq .9$).[9]

2.2.2. Learning time

Fig. 2 shows the mean learning time[10] for those learners who successfully learned each word, together with the learning time for an ideal cross-situational learner (e') and the expected learning time for the Random from $\mathscr{C}$ learning strategy. Looking only at those learners who successfully learned all words under a given mode of presentation, there is a significant effect of degree of referential uncertainty on learning time in both presentation modes (Consecutive: $N = 28$, $\chi^2_F(2) = 14.771$, $p = .001$, post hoc tests reveal a significant difference between $C = 2$ and $C = 8$, $z = 3.343$, corrected $p = .003$, with other pairwise comparisons being nonsignificant, $z \leq 2.081$, corrected $p \geq .063$; Interleaved: $\chi^2_F(2) = 9.579$, $p = .008$, post hocs reveal a significant difference between speed for $C = 2$ and $C = 5$, $z = 2.44$, corrected $p = .045$, with other pairwise comparisons n.s., $z \leq 2.16$, corrected $p \geq .093$).[11] To measure the effect of interleaving on learning speed for successful learners, we exploit both within-subjects and within-pairs analyses. The within-subjects analysis suggests that interleaving has no impact on learning speed for any value of C ($z \leq 1.343$, $p \geq .179$). However, the within-pairs analysis reveals that learning is significantly slower in the interleaved condition for $C = 5$ ($N = 27$, $z = 2.378$, $p = .017$) but not for $C = 2$ or $C = 8$ ($z \leq 1.164$, $p \geq .244$).

2.2.3. Learning strategies

A crucial part of our design, inspired by Gillette et al. (1999), was to gather an exposure-by-exposure indication of what participants think a word refers to, in order to see how each participant solves the task for each word. There are several ways to use cross-situational
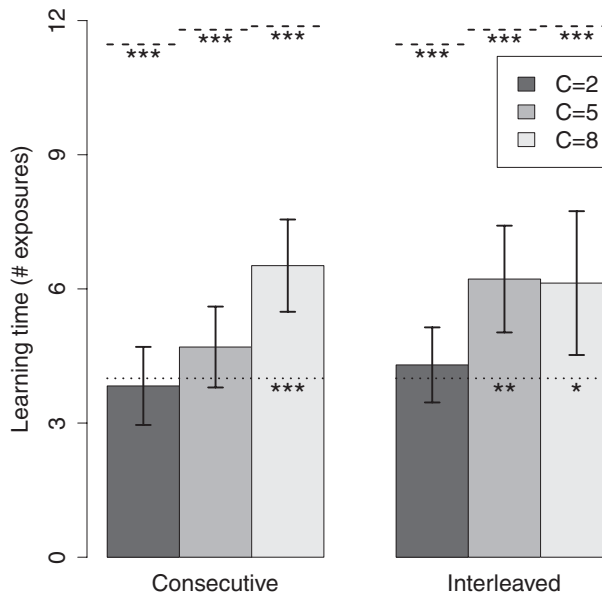
Fig. 2. Learning time for successful learners, compared with learning times for an ideal cross-situational learner (dotted lines) and the Random from $\mathscr{C}$ learner (dashed lines). Error bars give 95% confidence interval on the mean. Significant differences from the two baseline measures (according to one-sample Wilcoxon tests) are indicated by asterisks on the appropriate baseline (*, $p < .05$; **, $p < .01$; ***, $p < .001$).

information: Instead of the classic eliminative strategy, participants might select referents proportionately to the frequency with which they appear with a word; they might keep an initial guess about a word's meaning until disproved, or they might switch more readily. Based on an initial appraisal of data from a pilot study (see footnote 3), we identified four potential learning strategies[12]:

Random from $\mathscr{M}$: Select at random from the referents in the selection array, $\mathscr{M}$.

Random from $\mathscr{C}$: Select at random from the referents in the current context, $\mathscr{C}$.

Approximate XSL: If the referent chosen at the last exposure is in the current context, select it again; otherwise select from the referents in the current context, with a probability proportional to the frequency with which they have occurred in all exposures to this word.

Pure XSL: If the referent chosen at the last exposure is in the current context, select it again; otherwise select at random from the set of all referents that have occurred in every exposure to this word.

The latter two strategies make use of increasing degrees of cross-situational information. Both have a guess-and-test flavor, where participants keep choosing a previously chosen referent until its nonoccurrence in a context proves the choice incorrect, only then choosing a new referent. This seems (both impressionistically, and through our exploratory analysis) a broadly accurate characterization of how participants approached the task (although it is

possible that the repeated training-testing regime we used may itself have fostered this general approach).

How can we work out which strategy most closely matches participants' behavior? One possibility is to use performance on the task (learning success and learning time) to identify the strategy. However, preliminary analyses of the data suggest this is not a profitable approach, for two reasons. First, no single strategy adequately captures the population's performance with respect to learning success: The strategies outlined above predict success rates that are either substantially lower than those observed (the Random strategies) or substantially higher for the higher levels of *C* (the XSL strategies). Secondly, Pure XSL and Approximate XSL strategies make similar predictions with respect to learning time and are therefore indistinguishable, given our sample size. More generally, inferring the strategy from crude measures such as success rates or speed is difficult, particularly when different strategies make similar predictions.

A more fine-grained tool to fit behavioral data to learning strategies is needed; we therefore use the expectation maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977) to categorize each participant's behavior on each word. In essence, the EM algorithm identifies which of the four strategies above best describes the sequence of selections made by an experimental participant, and trades off both data fitting (strategy assignments which maximize the likelihood of the data are preferred) and overfitting (strategies which account for the behavior of few learners are dispreferred). Griffiths, Christian, and Kalish (2008) use a similar approach to distinguish experimental participants performing randomly from those performing in accordance with a nonrandom model. Ours is a minor complexification of their approach, as we seek to differentiate two kinds of random performance (Random from $\mathcal{M}$ and Random from $\mathcal{C}$) and two kinds of nonrandom performance (Approximate and Pure XSL).

2.2.3.1. Method for classifying behavior using expectation maximization: The likelihood of the sequence of selections *d* made by a participant on a word, given strategy *h*, is $P(d \mid h)$, where

$$P(d \mid h) = \prod_{i=1}^{e^{\max}} p(d_i = m \mid h, \ldots), \tag{1}$$

and where $p(d_i = m \mid h, \ldots)$ is the probability of selecting meaning *m* at exposure *i* given strategy *h* and the necessary elements of the exposure history *d* required by the strategy. The four strategies described above are formally defined as follows.

Random from $\mathcal{M}$ is defined simply as each meaning being chosen with an equal probability:

$$p(d_i = m \mid \text{Random from} \mathcal{M}) = \frac{1}{|\mathcal{M}|}, \tag{2}$$

where $\mathcal{M}$ is the set of referents in the selection array, and $|\mathcal{M}|$ its magnitude: In our experiment the selection array always contained 15 referents, hence $|\mathcal{M}| = 15$.

Random from $\mathscr{C}$ is similarly defined, but we allow a probability θ that a meaning not included in the context is selected in error:

$$p(d_i = m \mid \text{Random from } \mathscr{C}) = \begin{cases} (1-\theta)\frac{1}{|\mathscr{C}_i|}, \text{if } m \in \mathscr{C}_i \\ \theta \frac{1}{|\mathscr{M}|-|\mathscr{C}_i|}, \text{ otherwise} \end{cases} \tag{3}$$

where $\mathscr{C}_i$ is the set of referents in the context at exposure i, including the target, and $|\mathscr{C}_i|$ its magnitude.

Approximate XSL is the first strategy that integrates cross-situational information, and we therefore need to keep track of (a) previous choices, in particular the referent selected at the immediately preceding time step, d_{i-1}; (b) the frequency with which a given meaning m has occurred in $\mathscr{C}_i$ for all i exposures to date, which we denote $f_i(m)$:

$$p(d_i = m \mid \text{Approximate XSL}, d_{i-1} = m', f_i) = \begin{cases} (1-\theta), & \text{if } m' \in \mathscr{C}_i \quad \text{and } m = m' \\ (1-\theta)\frac{f_i(m)}{\sum_{m'' \in \mathscr{C}_i} f_i(m'')}, & \text{if } m' \notin \mathscr{C}_i \text{ and } m \in \mathscr{C}_i \\ \theta \frac{1}{|\mathscr{M}|-1}, & \text{if } m' \in \mathscr{C}_i \text{ and } m \neq m' \\ \theta \frac{1}{|\mathscr{M}|-|\mathscr{C}_i|}, & \text{if } m' \notin \mathscr{C}_i \text{ and } m \notin \mathscr{C}_i \end{cases} \tag{4}$$

The first two conditions cover the case where the strategy is applied correctly (occurring with probability $1 - \theta$). If the previous selection appears again in the current context (first case), it is maintained; otherwise (second case) a new selection is made from among the members of $\mathscr{C}_i$, weighted by the relative frequency with which these have occurred in $\mathscr{C}$ over the entire exposure history for this word. The final two conditions in Eq. (4) cover cases where the strategy is incorrectly applied: The previous selection is abandoned for a random choice despite it reappearing in the current context (third case), or a new selection is made from the complement of the current context (final case).

Pure XSL requires the learner to keep track of not only the immediately preceding selection but also the set of meanings that have occurred in $\mathscr{C}$ on every exposure for this word so far, $\mathscr{K}_i$:

$$p(d_i = m \mid \text{Pure XSL}, d_{i-1} = m', \mathscr{K}_i) = \begin{cases} (1 - \theta), & \text{if } m' \in \mathscr{K}_i \text{ and } m = m' \\ (1 - \theta)\frac{1}{|\mathscr{K}_i|}, & \text{if } m' \notin \mathscr{K}i \text{ and } m \in \mathscr{K}_i \\ \theta \frac{1}{|\mathscr{M}|-1}, & \text{if } m' \in \mathscr{K}_i \text{ and } m \neq m' \\ \theta \frac{1}{|\mathscr{M}|-|\mathscr{K}|}, & \text{if } m' \notin \mathscr{K}_i \text{ and } m \notin \mathscr{K}i \end{cases} \tag{5}$$

The first two cases in Eq. (5) again give the probabilities of a selection when the strategy is correctly applied (if the previous selection still appears in $\mathscr{K}$, it is maintained, otherwise

a new selection is made at random from $\mathcal{K}$), and the second two cases give the probabilities of selections when the strategy is deviated from.

In order to simplify the EM procedure, we assume that θ is the same for all strategies and all individuals but may vary according to referential uncertainty and mode of presentation. Let us assume that some proportion of the population $P(h)$ uses strategy h. For a given value of θ we can use Bayes' rule to compute the posterior probability that a participant i, producing data set D_i, is behaving according to strategy h:

$$P(h\,|\,D_i, \theta) = \frac{P(D_i\,|\,h, \theta)P(h)}{\sum_{h'} P(D_i\,|\,h', \theta)P(h')}, \tag{6}$$

where the sum is over all possible strategies—in our case, the four strategies defined above. Of course, the actual value of θ and the various priors are unknown. The EM algorithm provides a method for estimating these parameters, by iteratively re-estimating them, homing in on the set of parameters which maximizes the posterior probability of the data, using previous estimates of the parameters to calculate new best estimates and repeating until the estimates of the parameters stop changing. In more detail: We can use a previous estimate of θ and the various values of $P(h)$ to calculate the posterior probability distribution over strategies for each of our n participants (the Expectation step), and then use these quantities to re-estimate θ and $P(h)$ (the Maximization step) as follows (after Griffiths et al., 2008):

$$\widehat{P(h)} = \frac{\sum_{i=1}^{n} P(h\,|\,D_i, \theta)}{n} \tag{7}$$

$$\hat{\theta} = \underset{\theta}{\mathrm{argmax}} \sum_{i=1}^{n} \sum_{h} P(h\,|\,D_i, \theta) \log P(D_i\,|\,h, \theta). \tag{8}$$

The initial values of θ and $P(h)$ are arbitrary, and data for each level of referential uncertainty and mode of presentation are treated separately. We considered 999 values of $\hat{\theta}$ between 0 and 1, in increments of 0.001. We repeated the Expectation/Maximization loop until the parameters ceased to change, then selected the strategy with the maximum a posteriori (MAP) probability as the best characterization of each participant's behavior on each word.[13]

2.2.3.2. Results of the EM analysis: Table 3 shows the final values of the parameters provided by the EM analysis. The error parameter θ generally increases with C, and always has a higher value for interleaved presentations, as we might expect—following a strategy accurately is more difficult when there are interruptions to the sequence of exposures. While the estimated prior probability of the two Random strategies does not vary in any systematic way with C and mode of presentation, there appears to be a shift in prior probability from Pure XSL to Approximate XSL given higher C and/or interleaved presentation, which we discuss below.

Table 4 shows the results of the strategy classification (the MAP strategy for each participant for each word), confirming that participants do use cross-situational learning strategies

Table 3
Final error parameter (θ) and prior probabilities derived by the EM procedure

Presentation	C	θ	p(Random From $\mathscr{M}$)	p(Random From $\mathscr{C}$)	p(Approximate XSL)	p(Pure XSL)
Consecutive	2	0.038	0.000	0.138	0.191	0.670
	5	0.051	0.000	0.203	0.445	0.352
	8	0.055	0.000	0.117	0.665	0.219
Interleaved	2	0.057	0.000	0.127	0.651	0.221
	5	0.076	0.000	0.129	0.797	0.074
	8	0.060	0.000	0.210	0.790	0.000

Table 4
Distribution of the learning strategies used by experimental participants (out of 48)

Presentation	C	Random From $\mathscr{M}$	Random From $\mathscr{C}$	Approximate XSL	Pure XSL
Consecutive	2	0	6	3	39
	5	0	11	16	21
	8	0	6	31	11
Interleaved	2	0	6	42	0
	5	0	6	42	0
	8	0	10	38	0

in most cases, yet also suggesting that the type of strategy used depends on the level of referential uncertainty and mode of presentation. Focusing on the consecutive presentation data first, and dealing with participants who use a XSL strategy for all three words, we see a significant shift from Pure to Approximate XSL as C increases ($N = 34$, $Q(2) = 33.231$, $p < .001$, all post hoc tests significant, $Q(1) \geq 9.00$, $p \leq .009$ after Bonferroni correction). There is clearly no such shift mediated by C for interleaved presentation, as the EM analysis suggests that Pure XSL is not used for interleaved words. There is also a significant shift from Pure to Approximate XSL when comparing behavior on consecutive and interleaved presentation for all levels of C, indicated by an analysis for subjects who used a XSL strategy for both presentations ($C = 2$, $N = 39$, $Q(1) = 37.0$, $p < .001$; $C = 5$, $N = 34$, $Q(1) = 19.0, p < .001$; $C = 8, N = 36, Q(1) = 11.0, p = .001$).[14]

2.2.4. Summary of results

Our results show clear evidence of XSL, with better performance in all conditions than is achievable under the best-possible non-XSL strategy (Random from $\mathscr{C}$). We also show, in agreement with Yu and Smith (2007) and K. Smith et al. (2009), that word learning is significantly affected by the level of referential uncertainty: As C increases, success rates fall and learning times increase. Interleaving of exposures has a more marginal impact on learning success and speed, indicated by some analyses in some conditions. Finally, the strategic analysis suggests that participants use full-blown eliminative XSL as long as the task is reasonably easy, but switch to the less taxing Approximate XSL strategy when the demands of

the task increase (either through high *C*, or interleaved presentation). The effect of presentation mode on this shift is particularly marked: The EM analysis suggests that the true eliminative XSL strategy is never used when presentations are interleaved with presentations of other words. Humans are therefore capable of effective XSL, even under high referential uncertainty, but the rigor with which cross-situational information is exploited is modulated by the difficulty of the word learning task. Furthermore, the contrast between the large shift in learning strategy induced by interleaved presentation and the rather equivocal nature of the impact of interleaving on learning success and learning speed highlights how effective weaker, frequentist approximations to eliminative XSL can be.

3. Discussion

3.1. A continuum of learning strategies

Our strategy-based analysis implies that Pure XSL and Approximate XSL are distinct hypothesis-testing approaches to word learning. Yu, Smith, Klein, and Shiffrin (2007), however, argue that there is no fundamental difference between hypothesis testing and associative mechanisms. Similarly, we will argue here that there is a natural associative interpretation of our various strategies which illustrates the continuum on which they reside. In the context of the experiment outlined above, let us assume the following associative learning device:

1. Each possible word-meaning pairing is represented by a weighted association.
2. The occurrence of a meaning in the context associated with a target word increases the strength of that word-meaning association.
3. During testing, the device can select only from the meanings in the immediately preceding context (e.g., these associations are massively but temporarily boosted).
4. The device remembers its previous selection (again, perhaps the previous selection has its strength temporarily boosted).
5. During testing, the previous selection is simply repeated if present (following assumptions 3 and 4), or the meaning from the immediately preceding context with the highest activation (assumption 3) wins out as the best guess for the word meaning.

For such an associative device, a situation where activation levels perfectly reflect occurrence frequencies produces Pure XSL behavior: At each guess, only those meanings that have been present in every context to date will be selected (Yu et al., 2007). Now imagine that the association strengths are subject to noise—either they are subject to noisy updating, or they decay noisily, or the winner-take-all decision process is subject to error. Introducing noise means that the mapping between frequency and probability of selection becomes stochastic, yielding a range of strategies grading from pure XSL toward Approximate XSL (see Fig. 3). In the limit of noise, association strength becomes no cue to selection at all, and all meanings in the current context are equally probable. This strategy exploits the minimal amount of information across exposures, and can be described as follows:
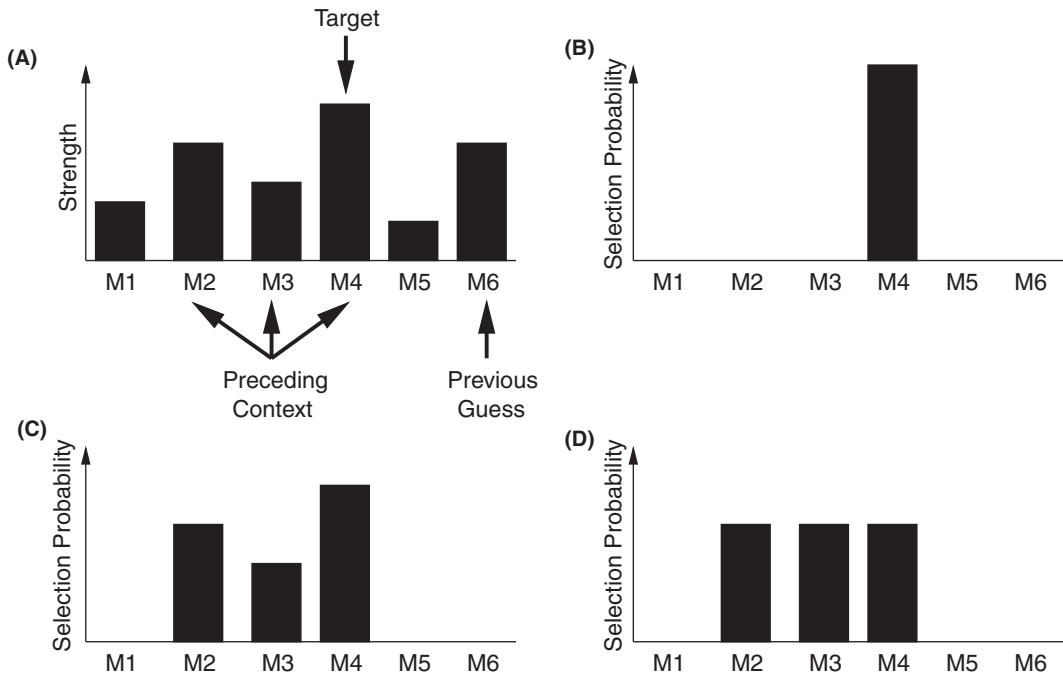
Fig. 3. A sketch of an associative instantiation of cross-situational learning. Strength of association between a single target word and a number of meanings (numbered M1–M6 here) are represented by height of vertical bars. (A) The frequency of co-occurrence yields a set of weighted associations. (B) When the association strengths are noise-free, the most frequently-occurring meaning is always selected; this is the Pure XSL strategy described in the text. (C) At intermediate levels of noise, a stochastic element is introduced into selection, with the probability of selection mirroring the underlying frequencies, as in the Approximate XSL strategy. (D) At high levels of noise, the frequency information stored in the association strengths is obscured, yielding Minimal XSL behavior.

Minimal XSL: If the referent chosen at the last exposure is in the current context, select it again; otherwise select at random from the referents in the current context.

All three cross-situational word learning strategies (Pure, Approximate, and Minimal XSL) could therefore be realized by a *single* associative learning device operating under different levels of noise. The experimental data suggests that, if we conceive of XSL in this way, increased referential uncertainty and interleaving of exposures lead to increased noise on association strengths. For example, high uncertainty may increase the likelihood that a learner will not notice elements of the context, thus introducing errors into the matrix of association strengths.

3.2. Strengths and weaknesses of our methodology

Our experimental method has several advantages. In common with Gillette et al. (1999), it allows an exposure-by-exposure insight into participants' hypothesis about what a target

word refers to. Unlike Gillette et al.'s more naturalistic stimuli, our artificial scenario permits control over the degree of referential uncertainty at each exposure, an attractive feature borrowed from Yu and Smith (2007)'s method. The most notable advantage of this approach is that it allows us to make a sensible guess about how our participants are tackling the XSL task, by fitting learning strategies to the data via the EM procedure.

The study does have a number of remaining weaknesses. The artificiality of the task means that the approach our participants adopt might bear little relation to how humans learn words in the real world. Future work will develop more naturalistic but equally controlled means of testing cross-situational word learning, perhaps involving context videos (following Gillette et al., 1999) with known levels of referential uncertainty (estimated as described in Blythe et al., 2010). Secondly, we are testing adults: Children may exhibit entirely different word learning behaviors. We would be particularly interested in whether child learners exhibit a similar shift in their use of cross-situational information as referential uncertainty increases, and whether this shift occurs at lower levels of uncertainty. Piccin and Waxman (2007) suggest that children are more likely to abandon successful guesses as to a word's meaning from exposure to exposure, suggesting that children's learning strategies may generally be characterized either by a higher θ parameter or the absence of a guess-and-test approach entirely. Finally, although our repeated testing approach provides the rich exposure-by-exposure data which constitutes a major strength of our method, it may also influence how participants approach the task, by fostering the guess-and-test approach which characterizes participants' behavior. More subtle methods (e.g., eye-tracking, as used in Yu & Smith, 2008) might allow estimates of learning strategies without explicitly probing a participant's hypothesis.

3.3. Implications for word learning in the real world

One possible interpretation of our results is that humans are powerful cross-situational learners, suggesting that word learning can be explained as a product of XSL, with minimal input from heuristics that reduce the referential uncertainty feeding into the XSL mechanism. However, we believe that these results, in conjunction with our analysis of lexicon learning times for cross-situational learners (Blythe et al., 2010), necessitate a more cautious conclusion at present.

Our experimental results indicate that adults apply weaker forms of XSL when referential uncertainty is high or exposures to a word do not occur consecutively. The real-world case is likely to be characterized by high uncertainty (extremely high uncertainty if we assume that word-learning heuristics are weak) and extensive interleaving (with substantially larger gaps between exposures than in our experiment). As such, we would expect XSL mechanisms applied to real word learning to make relatively minimal use of cross-situational information—in terms of the continuum of XSL strategies provided above, real-world XSL which is relatively unconstrained by word-learning heuristics might be best characterized as Minimal XSL. In Blythe et al. (2010), we estimate learning times for human-scale lexicons for Minimal, Approximate, and Pure XSL learners, and show that, for all strategies, lexicon learning time (number of exposures required to learn a set of

words) increases as referential uncertainty increases. However, weaker forms of XSL are disproportionately affected by increased referential uncertainty: As a function of referential uncertainty, lexicon learning time increases more rapidly for Minimal XSL than Approximate XSL, and more rapidly for Approximate XSL than Pure XSL. In combination with our experimental data, this indicates a double penalty for high referential uncertainty: Not only does higher referential uncertainty necessarily increase lexicon learning time, but it also induces a shift toward weaker forms of XSL, which increases lexicon learning time further. At some point, referential uncertainty will drive the required lexicon learning time beyond the amount of data that learners can expect to see. Quantifying this critical degree of referential uncertainty is problematic, as we do not yet know how the learning strategies adopted by learners changes under referential uncertainty higher than that explored here. We expect, however, that relatively unconstrained XSL will require learning times too high for human-scale lexicons: We anticipate that the battery of word-learning heuristics discussed in Section 1 is required to reduce referential uncertainty to relatively low levels (on the order of a few tens of possible word meanings per exposure) if the cross-situational learning of large lexicons is to be feasible.

4. Conclusion

We have demonstrated that cross-situational word learning is significantly affected by the level of referential uncertainty and the way in which words are presented: High referential uncertainty and interleaving of exposures lead to less successful, slower learning. Furthermore, we identify a continuum of possible cross-situational strategies and show that, although humans are effective cross-situational learners even at high levels of referential uncertainty, the rigor with which they exploit cross-situational information is modulated by the apparent difficulty of the task, as determined by degree of referential uncertainty and interleaving. Finally, we have shown that the variants of XSL described here can be explained in terms of a single underlying associative learning model.

Notes

1. Each word list contained the same number of words in each category. Word list 1: *oyb, cherve, fral, twilt, gotif, sladzene, midzivore, qualifor*; word list 2: *alk, benth, clow, smay, noblin, crigid, voonarist, fronarchy*.
2. We see significant practice effects in other XSL experiments (e.g., K. Smith et al., 2009), and as such generally include practice words to eliminate order effects from the data. While an anonymous reviewer rightly suggests that these practice effects are a promising area for future investigation, we do not address such questions here.
3. A pilot experiment ($N = 41$), exploring only consecutive presentation, produces similar results to the experiment described here with respect to learning success, learning time, and learning strategies applied.

4. This time limit was only reached once in the whole experiment.
5. In an interleaved exposure block, a participant received one exposure to the first word, followed by one exposure to a second word, followed by one exposure to a third word, followed by a second exposure to the first word and so on, with the C values of the first, second, and third words being determined by the ordering parameter.
6. Note that, in keeping the values of M, e^{max}, and e' constant, it is necessarily true that the distribution of frequencies with which distractors co-occur with the target varies with C: Controlling all these factors simultaneously is not possible.
7. There was no significant effect of any of the between-subjects factors on total learning success ($M = 4.54$ words learned out of six possible, $SD = 1.458$; no effect of word list, $z = 0.878$, $p = .38$; no effect of ordering of blocks, $z = 0.118$, $p = .906$; no effect of ordering of levels of C, $H(5) = 0.378$, $p = .996$), and results are therefore combined across orderings.
8. Such analyses are commonly used in time-to-event analyses, particularly in medical statistics. Our model was implemented using the coxph function in the survival package for the freely available statistical program R, which is published by R Foundation for Statistical Computing, Vienna, Austria. The term ''hazard'' derives from its use in clinical analyses, where the event in question is the emergence of a particular medical complication, or the death of the patient—in our model the event is the *word becoming learned*.
9. A nonparametric repeated-measures analysis (using Cochran's Q statistic) reveals a significant effect of level of referential uncertainty on success for both consecutive presentation ($Q(2) = 18.00$, $p < .001$ and interleaved presentation ($Q(2) = 29.83$, $p < .001$). However, mode of presentation in this analysis does not yield any significant effect on overall learning success for any value of C either in the within-subjects analysis ($C = 2$, $Q(1) = 2.00$, $p = .157$; $C = 5$, $Q(1) = 0.00$, $p = 1.0$; $C = 8$, $Q(1) = 2.67$, $p = .102$), or in an analysis within yoked pairs ($C = 2$, $Q(1) = 0.667$, $p = .414$; $C = 5$, $Q(1) = 0$, $p = 1$; $C = 8$, $Q(1) = 2.667$, $p = .102$). The mismatch between this analysis and the analysis presented in the main text speaks to the relatively weak impact of interleaved presentation on learning success.
10. Learning times are without exception non-normally distributed, necessitating the use of nonparametric statistics throughout. There was no effect of any between-participant factors (word list, block order, order of encountering levels of C) on average time taken to learn successfully learned words ($z \leq 0.588$, $H(5) = 7.73$, $p \geq .172$), and all results are therefore presented with all between-participants factors collapsed.
11. The yoked pair analysis also shows a significant effect of C on learning times in consecutive presentation ($N = 114$, $H(2) = 16.514$, $p < .001$) and a marginal effect with interleaved presentation ($N = 104$; $H(2) = 5.917$; $p = .052$), reflecting the reduced difference in learning times for $C = 5$ and $C = 8$ with interleaved presentation.
12. There are (infinitely) many strategies for exploiting cross-situational information, and it is likely that the actual strategies our participants used are not included in our

list of four possibilities. Nonetheless, our list is representative of the main classes of strategy that might be applied.

13. The mean posterior probabilities for all MAP strategies are above 0.75.

14. A within-pairs analysis for words learned cross-situationally in both presentation modes also yields significant effects for mode of presentation: $C = 2$, $N = 36$, $Q(1) = 34.0$, $p < .001$; $C = 5$, $N = 33$, $Q(1) = 19.00$, $p < .001$; $C = 8$, $N = 33$, $Q(1) = 9.00$, $p = .003$.

Acknowledgments

References

Akhtar, N., & Montague, L. (1999). Early lexical acquisition: The role of cross-situational learning. *First Language*, *19*, 347–358.

Baldwin, D. A. (1991). Infants' contribution to the achievement of joint reference. *Child Development*, *62*(5), 875–890.

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., et al. (2007). The English lexicon project. *Behavioral Research Methods*, *39*(3), 445–459.

Bloom, P. (2000). *How children learn the meanings of words*. Cambridge, MA: MIT Press.

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

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

Childers, J. B., & Pak, J. H. (2009). Korean- and English-speaking children use cross-situational information to learn novel predicate terms. *Journal of Child Language*, *36*, 201–224.

Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society Series B*, *34*(2), 187–220.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B*, *39*(1), 1–38.

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

Gillette, J., Gleitman, H., Gleitman, L., & Lederer, A. (1999). Human simulations of vocabulary learning. *Cognition*, *73*, 135–176.

Gleitman, L. R. (1990). The structural sources of verb meanings. *Language Acquisition*, *1*, 3–55.

Gleitman, L. R., Cassidy, K., Nappa, R., Papafragou, A., & Trueswell, J. C. (2005). Hard words. *Language Learning and Development*, *1*(1), 23–64.

Griffiths, T. L., Christian, B. R., & Kalish, M. L. (2008). Using category structures to test iterated learning as a method for indentifying inductive biases. *Cognitive Science*, *32*(1), 68–107.

Horst, J. S., & Samuelson, L. K. (2008). Fast mapping but poor retention by 24-month-old infants. *Infancy*, *13*(2), 128–157.

Macnamara, J. (1972). The cognitive basis of language learning in infants. *Psychological Review*, *79*, 1–13.

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

Nappa, R., Wessel, A., McEldoon, K. L., Gleitman, L. R., & Trueswell, J. C. (2009). Use of speaker's gaze and syntax in verb learning. *Language Learning and Development*, *5*(4), 203–234.

Piccin, T. B., & Waxman, S. R. (2007). Why nouns trump verbs in word learning: New evidence from children and adults in the human simulation paradigm. *Language Learning and Development*, *3*(4), 295–323.

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

Pinker, S. (1994). How could a child use verb syntax to learn verb semantics? *Lingua*, *92,* 377–410.

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

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

Smith, L. B., & Yu, C. (2008). Infants rapidly learn word-referent mappings via crosssituational statistics. *Cognition*, *106*(3), 1558–1568.

Smith, K., Smith, A. D. M., & Blythe, R. A. (2009). Reconsidering human cross-situational learning capacities: A revision to Yu and Smith's (2007) experimental paradigm. In N. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 2711–2716). Austin, TX: Cognitive Science Society.

Spruance, S. L., Reid, J. E., Grace, M., & Samore, M. (2004). Hazard ratio in clinical trials. *Antimicrobial Agents and Chemotherapy*, *48*(8), 2787–2792.

Tomasello, M., & Farrar, J. (1986). Joint attention and early language. *Child Development*, *57*, 1454–1463.

Xu, F., & Tenenbaum, J. B. (2007a). Sensitivity to sampling in Bayesian word learning. *Developmental Science*, *10*(3), 288–297.

Xu, F., & Tenenbaum, J. B. (2007b). Word learning as Bayesian inference. *Psychological Review*, *114*(2), 245–272.

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. B. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*, *18*(5), 414–420.

Yu, C., & Smith, L. B. (2008). What you learn is what you see: Using eye movements to study infant cross-situational word learning. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 1023–1028). Austin, TX: Cognitive Science Society.

Yu, C., Ballard, D. H., & Aslin, R. N. (2005). The role of embodied intention in early lexical acquisition. *Cognitive Science*, *29*, 961–1005.

Yu, C., Smith, L. B., Klein, K. A., & Shiffrin, R. M. (2007). Hypothesis testing and associative learning in cross-situational world learning: Are they one and the same? In D. S. McNamara & J. G. Trafton (Eds.), *Proceedings of the 29th Annual Conference of the Cognitive Science Society* (pp. 737–742). Austin, TX: Cognitive Science Society.