

Manuscript version: Author's Accepted Manuscript

The version presented in WRAP is the author's accepted manuscript and may differ from the published version or Version of Record.

Persistent WRAP URL:

<http://wrap.warwick.ac.uk/114370>

How to cite:

Please refer to published version for the most recent bibliographic citation information. If a published version is known of, the repository item page linked to above, will contain details on accessing it.

Copyright and reuse:

The Warwick Research Archive Portal (WRAP) makes this work by researchers of the University of Warwick available open access under the following conditions.

Copyright © and all moral rights to the version of the paper presented here belong to the individual author(s) and/or other copyright owners. To the extent reasonable and practicable the material made available in WRAP has been checked for eligibility before being made available.

Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge. Provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

Publisher's statement:

Please refer to the repository item page, publisher's statement section, for further information.

For more information, please contact the WRAP Team at: wrap@warwick.ac.uk.

Adults Use Distributional Statistics for Word Learning in a Conservative Way

Suzanne Aussems, and Paul Vogt

Abstract—This study examined how much adults rely on cross-situational information in word learning by comparing their gaze behavior in a word learning task with models of four learning strategies. We manipulated the input type of situations (consecutive vs. interleaved) and the co-occurrence frequencies between words and objects so that adult learners could infer correct word-object mappings based on cross-situational information. There are two key findings. First, an exposure-by-exposure analysis of gaze behavior during the word learning procedure revealed that most participants collected sufficient cross-situational information before they developed a preference for one particular word-object mapping, with consecutive as well as interleaved situations. Second, a classification approach in which individual gaze behavior was attributed to different word learning strategies showed that participants relied mostly on a Conservative cross-situational learning (XSL) strategy, compared to Associative XSL, Propose-but-Verify, and Random strategies. Adults relied on Conservative XSL when presented with consecutive and interleaved situations, but they shifted towards Associative XSL when presented with interleaved situations.

Keywords—Cross-situational word learning, Propose-but-Verify, adults, eye tracking, Expectation-Maximization algorithm

I. INTRODUCTION

Over the past decades, cross-situational learning (XSL) has played a prominent role in explaining the human ability to learn words [1]–[12]. In short, XSL is a cognitive mechanism that allows a learner to infer word-object mappings by collecting information from multiple situations. However, there is still much debate on how humans apply this mechanism. This study aims to provide converging evidence regarding the extent to which adults 1) use cross-situational information for word learning, 2) engage in guess-and-test behavior during word learning, and 3) rely on these strategies when words are presented in consecutive or interleaved situations.

XSL assumes that learners use distributional co-occurrence statistics of words and objects and consider only those word-object mappings that occur in most situations [7], [9]. To illustrate how XSL works, we use the Situations in Fig. 1. If a cross-situational learner tracks the co-occurrence frequencies of the novel word and objects across Situations, he could infer that the upper left object in Situation 1 is the object that occurs most frequently with the word *timilo* when he reaches Situation 4. There is abundant evidence that adults can, and indeed do, use cross-situational information for word learning [2]–[5], [8]–[12].

S. Aussems is with the Department of Psychology, University of Warwick, Coventry, UK, CV4 7AL, e-mail: s.aussems.1@warwick.ac.uk

P. Vogt is with the Department of Cognitive Science and Artificial Intelligence, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands

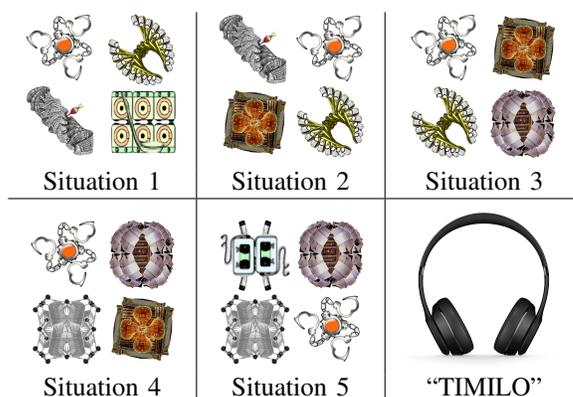


Fig. 1. Overview of a basic cross-situational word learning task. The five consecutive Situations display objects to which the word “TIMILO” could refer. Note that these Situations are usually presented to participants one at a time. The upper left object in Situation 1 appears most frequently across all five Situations. In fact, this object is the only object that has appeared in all Situations up to date in Situation 4. Using cross-situational information, a learner can thus infer that *timilo* refers to this object. Stimuli were developed by K. Smith, A. Smith, and Blythe [9].

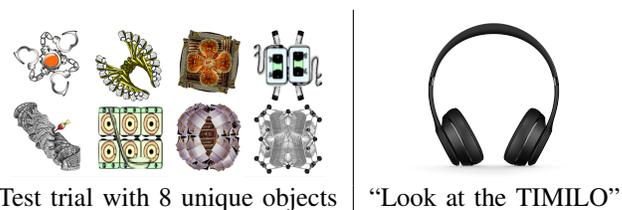


Fig. 2. Example of a test trial showing the eight unique objects from the consecutive Situations presented in Fig. 1.

A. Debate on Cross-Situational Learning

Researchers have argued that XSL cannot explain word learning under natural circumstances. For instance, Trueswell and colleagues suggested that storing the co-occurrence frequencies of multiple words and objects in human memory over a period time is so demanding, that XSL may only be used under greatly simplified circumstances, which are far simpler than those encountered by a language learner in a natural environment [13]. To account for this, they introduced a *Propose-but-Verify* strategy for word learning which requires a learner to track only one word-object mapping at a time [13]–[15]. According to the Propose-but-Verify (PbV) strategy, a learner quickly proposes a random word-object mapping when he first encounters a novel word, and he sticks to this mapping as long as possible. Considering the five Situations

depicted in Fig. 1, the PbV learner maps the novel word *timilo* randomly to one of the objects in Situation 1. If this object is still in Situation 2-5, the PbV learner sticks to this word-object mapping. However, if the hypothesized object is no longer consistent with previous Situations, the learner proposes another random word-object mapping without any recollection of the objects occurring in previous Situations. For instance, the right bottom object in Situation 1 is inconsistent with Situation 2. A learner who proposed this word-object mapping in Situation 1, will then select a random object in Situation 2, which could well be the newly introduced object on the bottom left. The PbV strategy requires a learner to track only one potential word-object mapping across situations, which is why this strategy is also called Minimal XSL [9].

B. Cross-Situational Learning or Single Hypothesis Testing?

Strikingly, in most of the studies that argue for a word learning strategy with a guess-and-test component, the model was derived from experiments in which participants were asked to select a word-object mapping in each situation. A forced-choice paradigm may thus require participants to engage in guess-and-test learning behavior. We present two lines of evidence to support this observation. First, the PbV strategy has been shown to explain word learning data obtained with a forced-choice word learning procedure. In the study by Trueswell and colleagues [13], adults were presented with a procedure in which they were required to pair visually presented sets of objects with target words on each exposure. As the sets of objects changed across situations for each word, the word-object mappings could be inferred from cross-situational information. Based on an analysis of the participants' clicking behavior during the word learning procedure, however, they found that their participants seemed to make random guesses when their latest proposal did not re-appear in the new situation, thus suggesting that they used the PbV strategy, rather than distributional co-occurrence frequencies of words and objects. Note that Trueswell et al. [13] also demonstrate PbV learning behavior with gaze data of adult learners, but these gaze data were collected during the forced-choice learning procedure, which may have influenced their gaze behavior (i.e., participants look where they click).

Second, integrated accounts of PbV and XSL have been shown to explain word learning data obtained with a forced-choice procedure. For instance, K. Smith and colleagues [9] designed a word learning task in which they presented adults with situations similar to those in Fig. 1. Their participants either viewed these situations consecutively for each word or interleaved, that is, the situations for one word were mixed with situations for other words. Participants were required to click on the object they thought the word referred to after each situation. Their clicking behavior was attributed to four different learning strategies (Pure XSL, Approximate XSL, and two random baseline strategies) using the Expectation-Maximization algorithm [16]. They found that participants adhered to Pure XSL with consecutive situations but shifted to Approximate XSL with interleaved situations. Similarly, Yurovsky and colleagues [11] presented adult participants with

a series of situations in which they heard a word, saw a number of objects, and were asked to guess the correct word-object mapping by clicking on the object. Their findings are explained by an integrated approach to XSL in which the learner uses both a guess-and-test approach and co-occurrence statistics of words and objects to identify correct word-object mappings. Following up on this, MacDonald, Yurovsky and Frank [17] demonstrated that especially when referential uncertainty was high, adult participants adhered to XSL, while they behaved more as single hypothesis testers when referential uncertainty was low. These studies suggest that the degree to which learners collect cross-situational information depends largely on the complexity of the task.

A plausible alternative for a forced-choice paradigm is a vision-based paradigm (e.g., passive look-and-listen or goal-based vision) [18] in which participants' gaze behavior provides an indication of their preference for an object, or multiple objects, when they hear a target word during the word learning procedure [12], [19]–[21]. In the study by Koehne and colleagues [20], participants were subjected to a passive look-and-listen paradigm. Participants' gaze behavior showed that they were able to distinguish between co-occurrence frequencies of the different (distractor) objects. However, this was only the case for objects that were actively selected during the word learning procedure, or for objects that received particular attention. Because co-occurrence frequencies were not necessarily all stored in memory, Koehne et al. [20] argue that there is a multiple-proposal account at play which corresponds to an extended PbV account for word learning rather than to XSL. However, in half of their experimental conditions participants selected the most likely object (based on cross-situational information) above chance level when they could not verify their previous guess. Thus, unlike a strict PbV theory would predict, learners can memorize more than their most recent choice. In the study by Koehne et al. [20], in which a vision-based paradigm was tested as an alternative to forced-choice, participants' gaze behavior was analyzed at test, but not during the training procedure. Based on the studies by Trueswell and colleagues [13], [20] it thus remains unclear whether learners use merely PbV or also rely on XSL.

It is plausible that learners could use more conservative learning strategies, in which they do not guess an initial word-object mapping, but track the distributional statistics of object appearances across situations in which a word is heard. This *Conservative XSL* strategy allows learners to gradually collect information to identify a correct word-object mapping, and to propose a word-object mapping only when sufficient cross-situational information has been obtained. This strategy is modeled after the Pure XSL strategy proposed by K. Smith and colleagues [9]. A conservative cross-situational learner will therefore not propose a word-object mapping until Situation 4 in Fig. 1, when he can infer that the upper left object in Situation 4 has appeared more often in the context of the word *timilo* than any of the other objects. To test our Conservative XSL strategy, we conducted a vision-based word learning experiment in which we tracked the gaze behavior of adult participants to investigate the development of word-object mappings in an exposure-by-exposure analysis. This

method allows us to attribute features key to the learning process of individual words and participants to Conservative XSL, Associative XSL, PbV, and a Random strategy.

C. The Present Study

To summarize, previous studies employed a forced-choice paradigm to investigate word learning strategies, which required participants to engage in a guess-and-test approach to word learning [9], [11], [13]. Therefore, it remains unclear to what extent human learners rely on PbV and cross-situational information for word learning. In order to investigate this, we use a vision-based paradigm, which allows participants to consider multiple word-object mappings during the word learning procedure without making a forced choice. We adapt the experimental design of K. Smith and colleagues [9] by recording participants' eye gaze during the learning procedure instead of their clicking behavior. We chose their design, because it includes input types in which situations for novel words are presented consecutively or interleaved with situations for other novel words. This allows us to investigate the contrasting claims with regard to the influence of consecutive and interleaved input types on word learning strategies. At the end of the consecutive and interleaved word learning procedures we measure participants' knowledge of the word-object mappings in test trials that show all unique objects from the training situations for a given word (see Fig. 2).

In our first analysis, we compare word learning performance and word learning speed (i.e. the number of exposures needed to identify the target object for a novel word) between training sessions that present adult learners with consecutive and interleaved situations (within-subjects manipulation). We predict participants will learn more words with consecutive than with interleaved situations. We also expect that participants will identify the target object earlier in the training procedure with consecutive than with interleaved situations.

Second, we analyze the gaze behavior during the training procedures in two ways: our first analysis is similar to the analysis of Trueswell and colleagues [13] and focuses on objects that participants prefer after the object they initially preferred disappears from the context; something we call a *preference-switch-analysis*. If participants switch to objects consistent with cross-situational information more often than to inconsistent objects in a subsequent situation, then this indicates they implicitly remembered objects from in previous situations. This would suggest that participants use cross-situational information for word learning rather than PbV.

Our third analysis is similar to the analysis of K. Smith and colleagues [9], who implemented the Expectation-Maximization algorithm to classify the clicking behavior of their participants during their word learning procedure based on models of different learning strategies. In our analysis, we use gaze behavior to estimate the likelihood of the sequence of object preferences for exposures to each word and per input type, given a Random strategy, PbV, Associative XSL, and our proposed Conservative XSL strategy in which participants do not focus on one particular object until they have collected sufficient information to disambiguate the target objects from the

distractors. Finally, we classify (combined) learning strategies per participant and input type of the situations (consecutive vs. interleaved) to see what type of strategies adult learners use, and under which memory demands. A consecutive presentation of situations requires participants to keep less objects in mind before the final situation of a word is shown than an interleaved presentation of situations. Therefore, an interleaved input type poses a higher demand on memory than a consecutive input type and we expect this to influence the learning behavior.

II. METHOD

A. Design

The experiment used a within-subjects design with input type (consecutive vs. interleaved) and situations (1-5) as the independent variables. We adapted the experimental design of K. Smith and colleagues [9] by reducing the number of situations for each word from twelve to five and by recording the gaze behavior of participants during the word learning procedure instead of their clicking behavior. Each participant learned four words with each input type of situations and there were eight test trials in total, four at the end of each input type. The dependent variable was the looking time towards target objects in the training situations and test trials.

B. Participants

We collected data from 92 Dutch native speakers (47 females, 45 males), all students in the Tilburg School of Humanities. Participants received course credit for participation. Ten participants were excluded from the analyses, because the eye tracker did not record their eye movements properly. The final sample consisted of 82 participants (45 females, 37 males) between 18–30 years old ($M = 22.46$, $SD = 2.72$).

C. Materials

Eight novel words were used to label novel objects, following Dutch phonotactical rules: *toekie*, *boezie*, *voolee*, *wootie*, *nieloo*, *wiepo*, *veegoo*, *reezoo*. Rounded and sharp syllables were balanced in each word to eliminate a potential influence from sound symbolism on learning behavior [22]. Audio samples of these words were generated using a female voice from a Dutch online text-to-speech generator available at <http://www.fluency.nl/international.htm>.

Next, 64 pictures of novel objects were randomly selected from the stimuli set developed by [9]. We created eight sets of eight objects for each word. One target object was randomly selected from each set and labeled with one of the novel words. The remaining stimuli in each set served as distractor objects. There was no overlap between sets to prevent participants from using knowledge about one word's referent to learn the correct referent of another novel word [23].

Subsequently, we created five situations for each word that included the target object and three distractor objects on a 2x2 display. Cross-situational information was manipulated by replacing one of the distractor objects systematically with one of the other distractor objects in each subsequent situation [13]. Fig. 1 shows an example set of five situations for the

word *timilo*, which illustrates how distractor objects were organized across situations. The eight sets of situations were then randomly assigned to the consecutive or interleaved input types of the word learning procedure.

In the consecutive condition, all five situations for one word were followed by all five situations for a second, third, and fourth word. In the interleaved condition, all first situations for four words were followed by all second, third, fourth, and fifth situations for these four words. Thus, in the interleaved condition, participants were shown four additional training trials, one for each of the other novel words, before they were shown the second training trial for the initial word. Test trials followed at the end of each condition and included the sets of the eight objects for each of the four words on a 2x4 display.

D. Apparatus

A SMI Vision RED 250 remote eye-tracking system was used for stimuli presentation and data collection. Stimuli were presented on a 22" computer screen via SMI Experiment Center 3.3 and gaze data from the eye-tracker were simultaneously collected via SMI iView X. Bright lights on both sides of the computer screen provided optimal calibration.

E. Procedure

Participants were tested individually in a soundproof booth in the lab. The distance between the participants and the computer screen was approximately 70 cm. The eye-tracker was calibrated for each participant using a 9-point calibration scheme. The experimenter validated if the estimation of the eye position was indeed close to the known calibration points. If errors occurred, the calibration session was repeated. After the experimenter left the booth, participants put on headphones and started the experiment. They were instructed via the screen to try and map a heard target word to the correct object, with the hint that the correct object was always displayed in the context of a heard word. Note that participants were not explicitly instructed to use a particular learning strategy.

Participants either completed the consecutive input type or the interleaved input type first. During the training procedure, participants were presented with each situation for 5000 ms, and all situations showed four objects while participants heard a target word. This duration was chosen to give participants enough time to scan four objects and was based on a small pilot study. Test procedures for the consecutive and interleaved input types followed when trials in each of the input types had finished. During the test procedures, participants were presented with the total set of objects for 8000 ms and they were instructed to focus on the objects that they thought corresponded to the played novel words. The duration in the test procedure was increased to allow participants to scan the eight objects shown (instead of 4 shown in the training phase).

We have made the audio files of the heard novel words available via Open Science Framework at osf.io/bskeh, and example video clips of the gaze behavior of one participant as it was monitored by the eye tracker during the task.

F. Data Analysis

Eye movements of the participants were analyzed using an Areas of Interest (AOI) approach. Equally-sized AOIs were drawn around all objects in training and test trials by a human coder. Looking times included fixations and saccades, which were assigned manually using the software and procedure by Cozijn [24]. Looking times for each object were normalized by dividing them by the total amount of looking time recorded per situation for each participant, and converted to percentages.

The objects participants looked at for more than 50% of the time during the test procedures were accepted as the chosen objects. Participants' looking times met this threshold in 94.8% of the test trials in the consecutive condition and 95.4% in the interleaved condition, suggesting that these generally reflected a choice when participants were instructed to "Look at the TIMILO". If this chosen object indeed corresponded to the target word, the word-object mapping was considered learned. The participants were scored one point for each learned word. If the chosen object did not correspond to the target word, or if none of the looking times met the >50% threshold, participants received zero points¹. Learning speed was operationalized as the percentage of time participants spent looking at target objects in each situation in the training procedures.

Word learning performance in the test trials (0=incorrect, 1=correct) following the consecutive and interleaved input types were analyzed with a mixed effects logistic regression analysis with input type as a fixed effect and participant and word as random effects.

Looking times (in percentages) for target objects during training were entered into a linear mixed-effects model that included input type and situation as fixed effects and participants and words as random effects.

All statistical analyses were carried out with the R software [25] using the *lme4* package [26]. Visual inspection of residual plots did not reveal any obvious deviations from homoscedasticity or normality. In all models, we started off with a maximal random effects structure including random slope and intercept variation, and the co-variation between the two, for participants and words [27]. We compared each model with updated versions of the model that systematically excluded each main effect and interaction term of interest using likelihood ratio tests (χ^2). The raw data files, R Markdown files, and code used for all the analyses and plots in this paper are available from the Open Science Framework at osf.io/bskeh.

Finally, we validated our >50% threshold for object preference during the training procedures by measuring the agreement between objects that met the threshold in the final situation for each word in each input type and the objects that met the threshold in the test procedures. In 82.9% of the

¹13 out of 656 data points (1.98%) did not meet our >50% threshold of the looking time and were treated as incorrect responses. The interpretation of our results is exactly the same with a threshold of >60% or >70% of the looking time. We checked these stricter thresholds because one could argue that looking times split between just two objects (49% vs. 51%) hardly indicate a preference for one of those objects. However, as their were four objects shown in each situation, differences smaller than 10% between objects that received the longest looking times occurred in less than 1% of our data.

final situations in the consecutive input type, and in 64.3% of the interleaved input type, participants showed a preference according to our threshold. Of these preferences, 83.1% of the preferred objects in the consecutive condition, and 75.4% in the interleaved condition, corresponded to objects that participants focused on in the test procedures. Our threshold thus identified objects that participants preferred at the end of the training procedures and in the test procedures.

G. Preference-Switch-Analysis

For the preference-switch-analysis, we first coded the situations of the training procedures in which one of the objects received more than 50% of the looking time. The threshold of more than 50% of the looking time for one object was chosen as it entails that participants could not have looked longer at any of the other objects. We extracted information about when a first object preference occurred across the five situations. Second, we coded whether the objects that participants preferred had appeared in all situations so far (i.e., whether the object was still in the potential set of objects that could refer to the target word). Third, we coded whether participants switched from one preference to another (e.g., from >50% looking time for object A to >50% looking time to object B in the next situation). Fourth, we coded switches to objects that had appeared in all situations so far. Finally, the objects to which participants switched were coded based on the situation in which they had first appeared, and based on how many subsequent instances they continued to appear. All these binary (1=switch, 0=no switch) dependent variables were entered into separate logistic regression analyses using models that controlled for a maximal random effects structure of participant and item variability where possible [27].

H. Expectation-Maximization Algorithm

The expectation-maximization (EM) algorithm [16] is an iterative classification method for finding the maximum likelihood or maximum a posteriori (MAP) estimates of parameters in statistical models, where the models depend on latent variables in addition to unknown parameters and known data observations. K. Smith and colleagues [9] used this method to estimate the likelihood that a particular word learning strategy explains the clicking behavior of their participants during the training procedure. We adapted their analysis to accommodate our experimental paradigm, which differs on two major accounts: 1) we analyzed sequences of gaze behavior instead of clicking behavior, and 2) our analysis only included looking times for objects presented in the context, whereas in the study by K. Smith and colleagues [9] participants were asked to also choose among potential objects that were not shown in the context, but appeared in a test trial after each situation.

Let d be the sequence of actions taken by a participant in subsequent situations of a given word, then the likelihood that, given strategy h , we observe this sequence is $P(d|h)$, where

$$P(d|h) = \prod_{i=1}^{t_{\max}} p(d_i|h, \epsilon, \dots), \quad (1)$$

in which $p(d_i|h, \epsilon, \dots)$ is the probability of behavior d_i at instance i , given strategy h , error parameter ϵ , and additional elements that describe the history of instances. We identify two typical gaze behaviors: 1) Participants' looking time for one object, $m_{\theta,i}$, in the context C_i exceeds our threshold of $\theta = 50\%$ of the looking time. We denote this behavior by $d_i = m_{\theta,i}$, where $m_{\theta,i} \in C_i$. 2) Participants do not look at an object for more than 50% of the time, but instead divide their looking time over all objects in the context C_i . We denote this behavior by $d_i \neq m_{\theta,i}$.

In our classification, we attribute sequences of gaze behavior to one of four learning strategies: a Random strategy, PbV, Associative XSL, and Conservative XSL. The probability functions of these strategies are below.

In calculating these probabilities, the error ϵ accounts for the probability that the observed behavior is inconsistent with the strategy under consideration. In other words, when calculating the probability for a specific learning strategy (e.g., Conservative XSL), the observed sequence of gaze behavior may not consistently follow the expected gaze behavior for this strategy, but there is a probability ϵ that this participant did use the given strategy, but with some error. For example, when the participant fixated on an object for longer than 50% of the time, this participant did not correctly apply the Conservative XSL strategy. However, this person would have correctly applied any of the other three strategies. The error ϵ accounts for the possibility that this participant would, overall, use the Conservative XSL strategy and not another strategy. The EM algorithm estimates the value of ϵ .

Random: is when a participant prefers to look at one object from the context C with an equal probability:

$$p(d_i|\text{Random}) = \begin{cases} (1 - \epsilon) \frac{1}{|C_i|}, & \text{if } d_i = m_{\theta,i} \\ \epsilon \frac{1}{|C_i|}, & \text{if } d_i \neq m_{\theta,i}, \end{cases} \quad (2)$$

where C_i is the set of objects in the context on instance i , and $|C_i|$ its size. In our experiment, we always presented participants with four objects, so $|C_i| = 4$. When a participant did not look at an object for more than 50% of the time, we assumed he did not apply the Random strategy. In this case, the probability is taken to be equal to ϵ times the a priori probability for looking at an object.

Propose-but-Verify (PbV): is the guess-and-test strategy proposed by [13] in which participants show an early preference, track the object that they preferred and look at this object again in case it re-appears in the context C_i . If participants' previous preference, $d_{i-1} = m_{\theta,i-1}$, does not re-appear in the present context, they randomly select a new object of preference. We attribute the following probability function to this strategy:

$$p(d_i|\text{PbV}) = \begin{cases} (1 - \epsilon), & \text{if } d_i = m_{\theta,i} \text{ and } m_{\theta,i} = m_{\theta,i-1} \\ (1 - \epsilon) \frac{1}{|C_i|}, & \text{if } d_i = m_{\theta,i} \text{ and } m_{\theta,i-1} \notin C_i \\ \epsilon \frac{1}{|C_i|-1}, & \text{if } d_i = m_{\theta,i}; m_{\theta,i-1} \in C_i \text{ and } m_{\theta,i} \neq m_{\theta,i-1} \\ \epsilon \frac{1}{|C_i|}, & \text{if } d_i \neq m_{\theta,i}. \end{cases} \quad (3)$$

The first two cases account for scenarios in which the strategy is correctly applied (with probability $1 - \epsilon$). If the object that a participant preferred on instance $i - 1$, $m_{\theta,i-1}$, occurs on instance i , the participant is expected to prefer that object

again (first case in Equation 3). If this object does not re-appear in the present context, C_i , then the participant is expected to prefer an object from the context with equal probability (second case). The other two cases describe scenarios in which PbV is incorrectly applied: The previously preferred object is in the context, but the participant prefers another object (third case) or no specific object (final case).

Associative XSL: is the guess-and-test XSL approach proposed by K. Smith and colleagues [9] and Yurovsky and colleagues [11]. In this approach, the participant needs to keep track of the frequencies $f_{m_{\theta,i}}$ with which mapping $m_{\theta,i}$ occurred in all i trials up to date. Following K. Smith and colleagues [9], we define the probabilities of the participant's gaze behavior as:

$$p(d_i|\text{Assoc. XSL}) = \begin{cases} (1-\epsilon), & \text{if } d_i = m_{\theta,i} \text{ and } m_{\theta,i} = m_{\theta,i-1} \\ (1-\epsilon) \frac{f_{m_{\theta,i}}}{\sum_{o \in C_i} f_{o,i}}, & \text{if } d_i = m_{\theta,i} \text{ and } m_{\theta,i-1} \notin C_i \\ \epsilon \frac{1}{|C_i|-1}, & \text{if } d_i = m_{\theta,i}, m_{\theta,i-1} \in C_i \text{ and } m_{\theta,i} \neq m_{\theta,i-1} \\ \epsilon \frac{1}{|C_i|}, & \text{if } d_i \neq m_{\theta,i}. \end{cases} \quad (4)$$

The first two cases, when the strategy is correctly applied, are the same as for PbV, but the assignment of the probability in the second case is now based on the co-occurrence frequencies of words and objects to fit with the Associative XSL account. So, when the previously proposed object is no longer in the context, we assume that participants prefer to look at each object proportionally to the frequency with which that object has occurred. The final two cases occur when Associative XSL is incorrectly applied. These conditions are the same as for PbV and we assume that the probabilities of looking at an object are the same as well.

Conservative XSL: is the strategy in which we assume that participants do not look at a particular object for more than 50% of the time until a single object can be identified as the target. This strategy is modeled after Pure XSL by K. Smith and colleagues [9]

$$p(d_i|\text{Cons. XSL}) = \begin{cases} (1-\epsilon) \frac{f_{m_{\max,i,i}}}{\sum_{o \in C_i} f_{o,i}}, & \text{if } d_i \neq m_{\theta,i} \text{ and } i \leq 3 \\ (1-\epsilon), & \text{if } d_i = m_{\theta,i}; m_{\theta,i} = m_{\max,i} \text{ and } i > 3 \\ \epsilon \frac{f_{m_{\theta,i}}}{\sum_{o \in C_i} f_{o,i}}, & \text{if } d_i = m_{\theta,i} \text{ and } i \leq 3; \text{ or} \\ \epsilon \frac{1}{|C_i|}, & \text{if } d_i = m_{\theta,i}; m_{\theta,i} \neq m_{\max,i} \text{ and } i > 3 \\ \epsilon \frac{1}{|C_i|}, & \text{if } d_i \neq m_{\theta,i} \text{ and } i > 3. \end{cases} \quad (5)$$

The first two cases describe scenarios in which participants applied Conservative XSL correctly. During the first three situations (first case), the target cannot be disambiguated from the distractors, and participants are not expected to look at one object for more than 50% of the time. Instead, they divide their attention over all the objects in the context. To be conservative in attributing a probability, we attribute a probability proportional to the frequency, $f_{m_{\max,i,i}}$, of that object on which participants fixated longest, i.e., $m_{\max,i} = \arg \max_m f_{m,i}$. The second case describes a scenario in which only one object (the target) has occurred in most contexts so far, and the participant has looked at this object. The final two cases describe the scenarios in which Conservative XSL was incorrectly applied. When the participant prefers an object during the first three situations (case 3a), he has developed a premature preference. The participant also does not apply Conservative XSL if he looks at a distractor during the fourth and/or fifth situation (case 3b). Finally, Conservative XSL is

incorrectly applied if the participant did not prefer an object when the target could be identified (in situation four and five). In all equations, we allowed participants to make an error with probability ϵ , which we assume is the same for all strategies, individual learners, and words, but may vary between the two input types. Still following [9], we applied Bayes' rule to calculate the posterior probability that an individual i generated the data D_i during the experiment by following strategy h , averaged over all W words the participant was exposed to, i.e.:

$$P(h|D_i, \epsilon) = \frac{1}{W} \sum_{n=1}^W \frac{P(D_{i,n}|h, \epsilon)P(h)}{\sum_{h'} P(D_{i,n}|h', \epsilon)P(h')}, \quad (6)$$

where $D_{i,n}$ is the data produced when learning word n , $P(h)$ is the prior probability that the participant used strategy h , and where the sum in the denominator is over all four strategies defined above. In this equation, the value of ϵ and the different priors are unknown, and we use the EM algorithm to estimate these values.

To obtain the best estimates of these parameters, we iteratively applied the EM algorithm to re-estimate these parameters until the parameters stopped changing more than a small value, δ . Following K. Smith and colleagues [9] and Griffiths and colleagues [28], each iteration consisted of two steps:

- 1) **Expectation step**: We used previous estimates of ϵ and $P(h)$ to calculate a posterior probability distribution over the four strategies that the N participants used, averaged over the four words per input type (Eq. 6).
- 2) **Maximization step**: We then used these values to re-estimate ϵ and $P(h)$ using:

$$\widehat{P(h)} = \frac{\sum_{i=1}^N P(h|D_i|\epsilon)}{N} \quad (7)$$

$$\widehat{\epsilon} = \arg \max_{\epsilon} \sum_{i=1}^N \sum_h P(h|D_i, \epsilon) \log P(D_i|h, \epsilon). \quad (8)$$

Initial values of ϵ and $P(h)$ are arbitrary. Following [9], we varied ϵ from 0 to 1 with increments of 0.001 to re-estimate ϵ as in Equation (8). The algorithm's loop was iterated until the differences between the new and old estimates for each parameter were smaller than $\delta = 0.001$. Finally, the strategy with the maximum a posteriori probability (MAP) was assigned to the gaze behavior for a given word, or participant.

III. RESULTS

A. General Findings

Table I shows the frequency distributions of the number of words identified correctly during the test procedures. Descriptively, participants learned all four words more often with consecutive than with interleaved situations.

We predicted the participants' accuracy of identifying word meanings in the test procedures with a `glmer` model that included input type as fixed effect and participant and word as random effects. Order of input type was originally included as a random effect, but dropped from the model because it had so little influence on the estimate of the fixed effect

TABLE I. FREQUENCY DISTRIBUTIONS OF THE NUMBER OF WORDS PARTICIPANTS LEARNED IN THE CONSECUTIVE AND INTERLEAVED INPUT TYPES OF THE WORD LEARNING PROCEDURE.

No. of words	No. of participants			
	Consecutive		Interleaved	
		%		%
0	1	1.2	1	1.2
1	1	1.2	13	15.9
2	10	12.2	22	26.8
3	22	26.8	22	26.8
4	48	58.6	24	29.3
Total	82	100.0	82	100.0

that it caused the model not to converge. The proportion of word meanings correctly identified in the test was significantly higher when the participants were trained with consecutive situations ($M = 0.85, SD = 0.36$) than with interleaved situations ($M = 0.67, SD = 0.47$), $\chi^2(1) = 10.62, p = .001$.

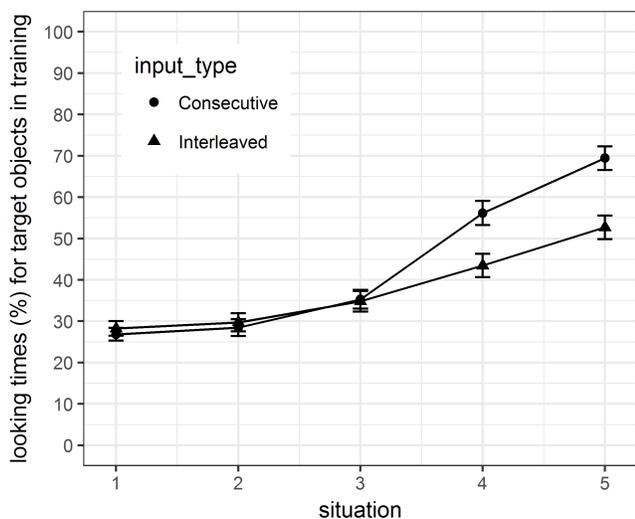


Fig. 3. Normalized looking times for target objects (in percentages on the y-axis) across situations (x-axis) in the consecutive and interleaved input types (shapes) of the training. Data are collapsed across four words in each input type. Error bars represent 95% confidence intervals of the means.

Fig. 3 shows the average time (%) that participants looked at the target objects across consecutive and interleaved situations. We predicted participants' looking times towards target objects in the training procedures using an `lmer` model that included input type and situation as fixed effects and participant and word as random effects. There was a significant interaction effect between input type and situation, on average looking times for target objects, $\chi^2(1) = 49.46, p < .001$. Bonferroni corrected post-hoc tests revealed no significant difference between the consecutive and interleaved input types in first, second, and third situations ($p > .05$). However, there was a significant difference between the two input types in fourth and fifth situations ($p < .001$), indicating that participants were more likely to look at target objects in these situations in the consecutive condition than in the interleaved condition.

B. Preference-Switch-Analysis

TABLE II. FREQUENCIES OF FIRST PREFERENCES (FP) AND SWITCHES MADE BY PARTICIPANTS BY INPUT TYPE AND SITUATION. IT IS SPECIFIED HOW OFTEN THESE FIRST PREFERENCES AND SWITCHES INVOLVED AN OBJECT THAT WAS IN THE SET OF MOST FREQUENTLY SHOWN OBJECTS IN ALL SITUATIONS UP TO DATE (K).

Situation	Consecutive				Interleaved			
	FP	In K	Switch	In K	FP	In K	Switch	In K
1	64	64	–	–	72	72	–	–
2	55	49	10	10	70	62	12	12
3	59	55	24	19	67	56	19	15
4	103	92	37	33	58	43	34	21
5	29	29	8	3	32	30	13	6
Total	310	289	79	65	299	263	78	54

Note. It was impossible to switch preferences in Situation 1, as this was the first exposure to a word.

In the first part of our preference-switch-analysis, we analyzed the preferences participants developed in the training procedures of the word learning task. In each input type (consecutive vs. interleaved), all 82 participants were trained on four different words, adding up to a total of 328 training sessions (i.e., sequences of five situations for a target word). Table II shows per input type in which situation of a training session a first preference was identified by the $>50\%$ threshold of looking time. In both input types, participants developed a first preference in the first situation in approximately 20% of all training sessions, indicating that participants generally did not start a training session by using a guess-and-test approach.

First preferences were entered into a `glmer` model with input type and situation as fixed effects and participant and word as random effects. We asked whether participants developed a preference during training, and if so, in which situation this first preference occurred. There was a significant interaction effect between input type and situation on first preferences, $\chi^2(4) = 20.21, p < .001$. Bonferroni-corrected post-hoc tests revealed that participants were more likely to show a first preference in fourth consecutive situations than in fourth interleaved situations ($p < .001$). Taking into account the numbers in Table II, this finding suggests that participants were more conservative in developing a first preference in consecutive situations than in interleaved situations. Every fourth situation provided participants with sufficient cross-situational information to distinguish the target objects from the distractors (see Fig 1). An analysis of first preferences for objects that appeared most frequently in all situations up to date, K , could tell us whether participants were more conservative with consecutive than with interleaved situations because they were collecting this information.

We first note that the vast majority of first preferences involved objects consistent with cross-situational information (93.2% in the consecutive condition and 88.0% in the interleaved condition). Again, an interaction effect between input type and situation was found when first preferences for objects in most contexts so far were entered into the analysis, $\chi^2(4) = 23.78, p < .001$. Bonferroni corrected post-hoc tests revealed that also for objects that appeared most frequently in all situations up to date, participants were more likely to

show a first preference in fourth consecutive situations than in fourth interleaved situations ($p < .001$). This finding suggests that participants collected more cross-situational information from consecutive situations than from interleaved situations.

In the second part of our analysis, we analyzed whether participants switched from preferring one object to another object in a subsequent situation according to the $>50\%$ threshold of looking time. Table II shows the number of training sessions in which participants made a switch in a given situation.

Switch data were entered into a `glmer` model analysis with input type and situation as fixed effects and participant and word as random effects. We asked whether participants switched preferences during training, and if so, in which situations these switches occurred. The main effect of situation on switches was significant, $\chi^2(4) = 122.36, p < .001$, but not the main effect of input type, $\chi^2(1) = 0.52, p = .469$, or the interaction, $\chi^2(4) = 2.78, p = .596$. Bonferroni-corrected post-hoc tests revealed that participants were more likely to switch in fourth situations than in any of the other situations ($p < .01$). Additionally, participants were less likely to switch in fifth situations than in third situations ($p = .039$). The fact that participants were more likely to switch in the fourth situation than in any of the other situations suggests that even learners who developed a premature preference used cross-situational information to switch preferences, because this is the situation in which they could have collected a sufficient amount of cross-situational information to distinguish the target objects from the distractors. An analysis of switches to objects that appeared most frequently in all situations up to date, K , could point out whether these participants were collecting cross-situational information.

The vast majority of switches, too, involved objects consistent with the available cross-situational information (86.5% for the consecutive condition and 77.6% for the interleaved condition). We found a similar pattern when switch data for objects that appeared most frequently in all situations up to date, K , were entered into the analysis. The main effect of situation on switches made to objects that appeared most frequently in all situations up to date was significant, $\chi^2(4) = 98.26, p < .001$, but not the main effect of input type, $\chi^2(1) = 0.01, p = .935$, or the interaction, $\chi^2(4) = 4.13, p = .389$. Bonferroni-corrected post-hoc tests revealed that participants were more likely to switch in fourth situations than in any other situations ($p < .05$). Additionally, participants were less likely to switch in fifth situations than in third situations ($p = .002$). The low number of switches made in the word learning procedure speaks to the relatively low number of participants that adhered to a guess-and-test strategy for word learning. So far, the findings suggest that even those participants who adhered to a guess-and-test strategy used cross-situational information during training.

In the final part of our preference-switch-analysis, we analyzed the objects that participants preferred after they made a switch. We asked whether participants were more likely to switch to target objects than to distractor objects. Table III shows the frequency with which participants switched to the different objects per input type.

Switches to different objects were entered into a `glmer`

TABLE III. SWITCH FREQUENCIES FOR OBJECTS SHOWN DURING THE TRAINING SESSIONS.

Object	Consecutive	Interleaved
	N	N
1 = target	46	41
2 = distractor in Situation 1, 2 and 3	15	11
3 = distractor in Situation 1 and 2	4	2
4 = distractor in Situation 1	–	–
5 = distractor in Situation 2, 3 and 4	0	0
6 = distractor in Situation 3, 4 and 5	9	14
7 = distractor in Situation 4 and 5	5	10
8 = distractor in Situation 5	0	0
Total	79	78

Note. It was impossible to switch to Object 4 as it appeared only in Situation 1.

model with input type and object as fixed effects and participant as random effect. The main effect of object on switches was significant, $\chi^2(1) = 308.99, p < .001$, but not the main effect of input type, $\chi^2(1) = 0.00, p = .999$, or the interaction, $\chi^2(7) = 5.17, p = .639$. Bonferroni-corrected post-hoc tests indicated that participants were generally more likely to switch to target objects than to any of the distractors ($p < .001$ for all comparisons).

C. Classification of Learning Strategies Based on Gaze

Table IV summarizes the findings of the EM algorithm, which we implemented to classify learning strategies per training session for each word (i.e., each sequence of five situations for a target word). In both the consecutive and interleaved input types, none of the participants used the Random strategy or the PbV strategy. In the consecutive input type, the classifier identified that participants used the Associative XSL for 8.54% of the 328 training sessions and Conservative XSL for 91.46% of the training sessions. In the interleaved input type, participants used the Associative XSL for 14.49% of the sessions and Conservative XSL for 85.06% of the sessions.

TABLE IV. CLASSIFICATION OF STRATEGIES BY WORD AND INPUT TYPE WITH THE EXPECTATION-MAXIMIZATION ALGORITHM.

	Consecutive		Interleaved	
	Prior	$N(\%)$	Prior	$N(\%)$
	$\epsilon = .097$		$\epsilon = .175$	
Random	.000	0 (0.00)	.000	0 (0.00)
Propose-but-Verify	.013	0 (0.00)	.055	0 (0.00)
Associative XSL	.143	28 (8.54)	.381	49 (14.94)
Conservative XSL	.844	300 (91.46)	.565	279 (85.06)
min MAP	.608		.493	
avg MAP	.968		.923	
Total		328 (100.0)		328 (100.0)

Looking at the strategies that individual participants used, we found that the vast majority of participants relied almost exclusively on Conservative XSL in both input types (see Table V). However, participants did not always use one specific learning strategy for all the words. In the consecutive condition, only two participants relied solely on Associative XSL and 12 participants used both Associative XSL and Conservative XSL. In the interleaved condition, three participants relied

exclusively on Associative XSL and 23 participants relied on both Associative XSL and Conservative XSL.

TABLE V. CLASSIFICATION OF (COMBINED) LEARNING STRATEGIES PER PARTICIPANT AND INPUT TYPE USING THE EXPECTATION-MAXIMIZATION ALGORITHM.

	Consecutive <i>N</i> (%)	Interleaved <i>N</i> (%)
Associative XSL	2 (2.44)	3 (3.66)
Conservative XSL	68 (82.93)	56 (68.29)
Associative XSL & Conservative XSL	12 (14.63)	23 (28.05)
Total	82 (100.0)	82 (100.0)

Note. The Random and PbV strategies were omitted from the table, since the EM did not find any occurrences of them.

IV. DISCUSSION

In this paper, we investigated how much adults rely on cross-situational information for word learning by comparing their gaze behavior with models of four learning strategies. To this aim, we designed an experiment in which participants learned word-object mappings in training sessions with both consecutive and interleaved situations, while we tracked their eye movements—a method that has proven adequate for studying this kind of learning behavior [29]. Our main questions are: To what extent do adults 1) use cross-situational information for word learning, 2) apply a guess-and-test strategy during learning, and 3) rely on these strategies when words are presented in consecutive or interleaved sequences of situations. To summarize, our analyses indicate that some adult learners used a guess-and-test strategy for word learning, but generally they relied on distributional co-occurrence frequencies of words and objects in the word learning task, and they were conservative in developing a preference for a particular word-object mapping. Adults showed to use Conservative XSL when presented with consecutive and interleaved situations, but they shifted towards Associative XSL when presented with interleaved situations.

These conclusions are based on three analyses. Our first analysis showed that participants learned words in our experiment well above chance in both conditions, but they learned more words with consecutive situations than with interleaved situations. This finding is consistent with findings from K. Smith and colleagues [9], who found that fewer participants could learn all words with interleaved situations than with consecutive situations, and that learning took them longer with interleaved than with consecutive situations. Our analysis of gaze behavior for target objects during the training procedure clearly shows that participants appeared more confident about the correct word-object mapping after the third situation in the consecutive condition than in the interleaved condition. This finding suggests that participants learned words faster in the consecutive condition than in the interleaved condition in our experiment too. The difference between the consecutive and interleaved input types is best explained by the fact that cross-situational information was presented sequentially, one word at a time, in the consecutive condition, but sequentially in parallel for all words simultaneously in the interleaved condition, thus demanding less memory resources.

Second, our preference-switch analysis demonstrates that participants often did not show a preference for an object before they had collected sufficient cross-situational information to disambiguate the target object from the distractors. Hence, participants generally did not use a guess-and-test strategy for word learning. In both the consecutive and interleaved input types, participants revealed a preference for an object in only 20% of the first situations. In the consecutive input type, this number gradually increased until situations in which participants could have collected sufficient information to disambiguate the target objects. Participants developed premature guesses more often in the interleaved input type, but not necessarily in first situations. Thus, participants seem to use more conservative learning strategies than accounts with a guess-and-test component would predict. This is inconsistent with findings from studies that employed a forced-choice paradigm in which this conservative behavior could not be detected [5], [9], [11], [13].

Additionally, the preference-switch analysis showed that participants who switched preferences did not switch to a random object as PbV would predict, but predominantly switched to an object that was most frequently presented in the situations up to date. Moreover, participants were most likely to switch preferences at the exact time in the procedure where they could have distinguished the target objects from the distractors. Furthermore, participants preferred target objects over any of the distractor objects when they switched preferences, regardless of whether they were presented with consecutive or interleaved situations. In contrast to Trueswell and colleagues [13], our analyses thus show that participants have implicit memory of objects that have been presented frequently in previous situations. This additional finding suggests that adults in our experiment tracked the co-occurrence frequencies of words and objects during the word learning procedure in both the consecutive and interleaved input types. Findings from our preference-switch analysis are generally consistent with a XSL account for word learning. In cases where participants developed a premature guess for a word-object mapping, they updated this guess with cross-situational information, which corresponds to XSL rather than PbV.

Third, we used the expectation-maximization algorithm to attribute gaze behavior during the training procedure to different word learning strategies. The EM algorithm attributed most gaze behaviors to Associative XSL and Conservative XSL. Strikingly, none of the gaze behaviors were attributed to PbV. The vast majority of gaze behaviors were attributed to Conservative XSL and participants appeared to rely on this strategy more strongly in the consecutive than interleaved condition, which is consistent with the findings from our preference-switch analysis. This is also in line with findings from [9], who found that participants switched to a less demanding XSL strategy when presented with interleaved situations.

The finding that our classifier identified the use of both Associative XSL and Conservative XSL is consistent with findings from K. Smith and colleagues [9], however they did not include Conservative XSL in their classification, but Pure XSL. Their EM analysis showed that participants relied on both Associative XSL and Pure XSL in the consecutive

condition and that they switched to Associative XSL in the interleaved condition. In addition, their classifier identified that some of their participants used a Random strategy. We think that this may indicate an occasional reliance on PbV, but K. Smith and colleagues [9] did not consider PbV (Minimal XSL in their study) to be a viable strategy and therefore excluded it from their classification analysis.

Some participants used a combination of word learning strategies, which is consistent with the hybrid XSL account developed by Yurovsky and colleagues [11], [17]. Their studies showed that learners tend to use a guess-and-test strategy similar to PbV when presented with a low number of distractors, but that they keep track of the distributional co-occurrence statistics of words and objects when they are presented with a high number of distractors, similar to Conservative XSL and Associative XSL. They argue that the underlying cognitive model is the same, but the differences they found were due to graded differences in memory and attention constraints. Yurovsky and colleagues [11], [17] manipulated the number of distractors shown in situations, but we manipulated the input type of situations in our experiment following K. Smith and colleagues [9]. Based on our experiment, a similar argument can be made that in the consecutive condition, which is less complex, learners rely more on Conservative XSL than in the more complex interleaved condition.

Trueswell and colleagues [13] suggested that XSL may be only be a plausible strategy for word learning under greatly simplified circumstances, and that human learners would rely on PbV with more natural, interleaved instances. Our findings suggest that participants who rely on a guess-and-test strategy in the interleaved condition, use cross-situational information to update their word-object mappings rather than PbV. XSL thus seems a plausible strategy even for learning with interleaved situations.

Our analyses are based on gaze behavior, which we recorded using a vision-based paradigm. We argue that the reason we did not observe guess-and-test learning behaviors as often as in previous research [13]–[15] is because participants were not forced to make a choice during the training procedure. Our finding that participants can learn words using cross-situational information in this way are in line with other studies that used a vision-based paradigm [12], [19], [21].

One could argue that a vision-based paradigm does not measure whether or not people developed a preference for an object. However, other studies have validated this approach and consider it reliable [12], [19]–[21]. We assumed that participants developed a preference for an object when they looked at that object for more than 50% of the time during a situation in the training procedure. Our validation of this approach revealed that participants generally chose objects in the test procedure which received more than 50% of their looking time at the end of the training procedure. This indicates that our interpretations of the findings are valid within the scope of this experiment.

It is, of course, important to ask how these findings generalize to more realistic word learning contexts, where referential uncertainty may be much more complex and where target referents may not be present in the hear-and-now. It is

theoretically implausible that XSL is the sole learning mechanism under continuously high levels of referential uncertainty [30]. Referential uncertainty must be reduced, for instance by applying heuristics using social cues (e.g., eye gaze, pointing), cognitive and pragmatic constraints (e.g., whole object bias, mutual exclusivity, principle of contrast), and sentential constraints [31]–[34]. However, such heuristic can be considered as mechanisms to reduce referential uncertainty, after which XSL learning can be applied [35]. When uncertainty becomes very low or when a learner is forced to make a choice, he or she may use a guess-and-test strategy (cf. [17]). Otherwise, the Conservative XSL strategy is a likely candidate.

One of the assumptions of XSL is that the referents of words are present in the contexts in which these words are heard. Learners in real life, however, will come to learn that sometimes the referent of a word is not present. This kind of noise hampers XSL, but various computational studies (e.g., [7], [36]) have shown that XSL can deal with this, as long as this occurs occasionally. If it would occur regularly, then additional heuristics would become essential, but it would still not invalidate XSL as an underlying learning mechanism that can be used. Pointing gestures can help narrow down the context of possible word referents, even when the referent of a word is absent. For instance, abstract deictic gestures [37] indicate seemingly empty locations in the gesture space to refer to what used to be there or to refer back to entities that were temporarily assigned this location during discourse.

It is also important to discuss how our findings generalize to young word learners. There is abundant evidence that children can and do use XSL [38]–[40], but it is yet unclear whether and to what extent they would apply Conservative XSL or a guess-and-test strategy. Some of the work by L. Smith and colleagues suggest that, especially for toddlers, parents naturally label a novel object when this object is in the child's view, which reduces referential uncertainty considerably [41]. Similarly, when young children point to entities in their direct environment, parents often label the things that children are already paying attention to, which also considerably narrows down the context of possible referents for spoken words [42]. Future research could investigate to what extent Conservative XSL or guess-and-test strategies interact with these social learning mechanisms.

V. CONCLUSION

To conclude, our study shows that adult learners use cross-situational information for word learning. They are rather conservative in developing preferences for particular word-object mappings, and do not often engage in guess-and-test behavior. Adults use cross-situational information regardless of whether word learning situations are presented consecutively or interleaved with situations for other words. However, they tend to shift from Conservative XSL to Associative XSL when they are presented with interleaved situations.

REFERENCES

- [1] C. Fisher, D. Hall, S. Rakowitz, and L. Gleitman, "When it is better to receive than to give: Syntactic and conceptual constraints on vocabulary growth." *Lingua*, vol. 92, pp. 333–375, 1994.

- [2] J. Gillette, L. Gleitman, H. Gleitman, and A. Lederer, "Human simulations of vocabulary learning." *Cognition*, vol. 73, no. 2, pp. 135–176, 1999.
- [3] G. Kachergis and C. Yu, "Continuous measure of word learning supports associative model." in *Joint IEEE International Conferences on Development and Learning and Epigenetic Robotics*, 2014.
- [4] —, "Observing and modeling developing knowledge and uncertainty during cross-situational word learning." *IEEE Transactions on Cognitive and Developmental Systems*, vol. 10, pp. 227–236, 2017.
- [5] B. McMurray, J. Horst, and L. Samuelson, "Word learning as the interaction of online referent selection and slow associative learning." *Psychological Review*, vol. 119, pp. 831–877, 2012.
- [6] S. Pinker, *Learnability and cognition: The acquisition of argument structure*. The MIT Press, 1989.
- [7] J. Siskind, "A computational study of cross-situational techniques for learning word-to-meaning mappings." *Cognition*, vol. 61, no. 1-2, pp. 39–91, 1996.
- [8] K. Smith, A. Smith, and R. Blythe, "Reconsidering human cross-situational learning capacities: A revision to Yu & Smith's (2007) experimental paradigm." in *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, N. Taatgen and H. van Rijn, Eds. Austin, TX: Cognitive Science Society, 2009, pp. 2711–2716.
- [9] —, "Cross-situational learning: an experimental study of word-learning mechanisms." *Cognitive Science*, vol. 35, no. 3, pp. 480–498, 2011.
- [10] D. Yurovsky, C. Yu, and L. Smith, "Competitive processes in cross-situational word learning." *Cognitive Science*, vol. 37, no. 5, pp. 891–921, 2013.
- [11] D. Yurovsky and M. Frank, "An integrative account of constraints on cross-situational learning." *Cognition*, vol. 145, pp. 53–62, 2015.
- [12] C. Yu and L. Smith, "Rapid word learning under uncertainty via cross-situational statistics." *Psychological Science*, vol. 18, no. 5, pp. 414–420, 2007.
- [13] J. Trueswell, T. Medina, A. Hafri, and L. Gleitman, "Propose but verify: Fast mapping meets cross-situational word learning." *Cognitive Psychology*, vol. 66, no. 1, pp. 126–156, 2013.
- [14] T. Medina, J. Snedeker, J. Trueswell, and L. Gleitman, "How words can and cannot be learned by observation." *PNAS*, vol. 108, no. 22, pp. 9014–9019, 2011.
- [15] K. Woodard, L. Gleitman, and J. Trueswell, "Two- and three-year-olds track a single meaning during word learning: Evidence for propose-but-verify." *Language Learning and Development*, vol. 12, no. 3, pp. 252–261, 2016.
- [16] A. Dempster, N. Laird, and D. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society Series B*, vol. 39, no. 1, pp. 1–38, 1977.
- [17] K. MacDonald, D. Yurovsky, and M. C. Frank, "Social cues modulate the representations underlying cross-situational learning," *Cognitive psychology*, vol. 94, pp. 67–84, 2017.
- [18] A. P. Salverda, M. Brown, and M. K. Tanenhaus, "A goal-based perspective on eye movements in visual world studies," *Acta psychologica*, vol. 137, no. 2, pp. 172–180, 2011.
- [19] S. Fitneva and M. Christiansen, "Looking in the wrong direction correlates with more accurate word learning." *Cognitive Science*, vol. 35, no. 2, pp. 367–380, 2011.
- [20] J. Koehne, J. Trueswell, and L. Gleitman, "Multiple Proposal Memory in Observational Word Learning," in *Proceedings of the 35th Annual Meeting of the Cognitive Science Society*, M. Knauff, M. Pauen, N. Sebanz, and I. Wachsmuth, Eds. Austin, TX: Cognitive Science Society, 2013, pp. 805–810.
- [21] H. Vlach and S. Johnson, "Memory constraints on infants' cross-situational statistical learning." *Cognition*, vol. 127, no. 3, pp. 375–382, 2013.
- [22] M. Imai and S. Kita, "The sound symbolism bootstrapping hypothesis for language acquisition and language evolution," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 369, no. 1651, p. 2013028, 2014.
- [23] E. Markman and G. Wachtel, "Children's use of mutual exclusivity to constrain the meanings of words." *Cognitive Psychology*, vol. 20, no. 2, pp. 121–157, 1988.
- [24] R. Cozijn, "Het gebruik van oogbewegingen in leesonderzoek." *Tijdschrift voor taalbeheersing*, vol. 28, no. 3, pp. 220–232, 2006.
- [25] R Development Core Team, *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2011, ISBN 3-900051-07-0. [Online]. Available: <http://www.R-project.org>
- [26] D. Bates, M. Mächeler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4." *Journal of Statistical Software*, vol. 67, no. 57, pp. 1–48, 2015.
- [27] D. Barr, R. Levy, C. Scheepers, and H. Tily, "Random effects structure for confirmatory hypothesis testing: Keep it maximal," *Journal of Memory and Language*, vol. 68, no. 3, pp. 255–278, 2013.
- [28] T. Griffiths, B. Christian, and M. Kalish, "Using category structures to test iterated learning as a method for identifying inductive biases," *Cognitive Science*, vol. 32, no. 1, pp. 68–107, 2008.
- [29] C. Yu and L. Smith, "What you learn is what you see: using eye movements to study infant cross-situational word learning." *Developmental Science*, vol. 14, no. 2, pp. 165–180, 2011.
- [30] P. Vogt, "Exploring the robustness of cross-situational learning under Zipfian distributions." *Cognitive Science*, vol. 36, no. 4, pp. 726–739, 2012.
- [31] E. V. Clark, *The lexicon in acquisition*. Cambridge University Press, 1993.
- [32] G. Hollich, K. Hirsh-Pasek, and R. Golinkoff, "Breaking the language barrier: An emergentist coalition model for the origins of word learning." *Monographs of the Society for Research in Child Development*, vol. 65, no. 3, pp. i–vi, 1–123, 2000.
- [33] J. Macnamara, *Names for things: a study of human learning*. Cambridge, MA: MIT Press, 1982.
- [34] J. Koehne and M. W. Crocker, "The interplay of cross-situational word learning and sentence-level constraints," *Cognitive science*, vol. 39, no. 5, pp. 849–889, 2015.
- [35] P. Vogt and J. Mastin, "Rural and urban differences in language socialization and early vocabulary development in mozambique." in *Proceedings of the 35th Annual Conference of the Cognitive Science Society*, M. Knauff, M. Pauen, N. Sebanz, and I. Wachsmuth, Eds. Austin, TX: The Cognitive Science Society, 2013, pp. 3787–3792.
- [36] J. De Beule, B. De Vylder, and T. Belpaeme, "A cross-situational learning algorithm for damping homonymy in the guessing game," in *ALIFE X. Tenth International Conference on the Simulation and Synthesis of Living Systems*, L. Rocha, L. Yaeger, M. Bedau, D. Floreano, R. Goldstone, and A. Vespignani, Eds. Cambridge, MA: MIT Press, 2006.
- [37] D. McNeill, *Hand and Mind: What Gestures Reveal about Thought*. University of Chicago Press, 1992.
- [38] N. Akhtar and L. Montague, "Early lexical acquisition: the role of cross-situational learning." *First Language*, vol. 19, no. 57, pp. 347–358, 1999.
- [39] J. Childers and J. Pak, "Korean- and English-speaking children use cross-situational information to learn novel predicate terms." *Journal of Child Language*, vol. 36, no. 1, pp. 201–224, 2009.
- [40] L. Smith and C. Yu, "Infants rapidly learn word-referent mappings via cross-situational statistics." *Cognition*, vol. 106, no. 3, pp. 1558–1568, 2008.
- [41] H. Yoshida and L. B. Smith, "What's in view for toddlers? using a head camera to study visual experience." *Infancy*, vol. 13, pp. 229–248, 2008.
- [42] J. Iverson and S. Goldin-Meadow, "Gesture paves the way for language development," *Psychological Science*, vol. 16, pp. 367–371, 2005.