Computational evidence for effects of memory decay, familiarity preference and mutual exclusivity in cross-situational learning

Heikki Rasilo (heikki.rasilo@aalto.fi)
Aalto University, Dept. Signal Processing and Acoustics, Otakaari 5 A, FI-02150 Espoo, Finland &
Vrije Universiteit Brussel, Artificial Intelligence Laboratory, Pleinlaan 2, 1050 Elsene, Belgium

Okko Räsänen (okko.rasanen@aalto.fi)
Aalto University, Dept. Signal Processing and Acoustics, Otakaari 5 A, FI-02150 Espoo, Finland

Abstract

Human infants learn meanings for words in interaction with their environment. Individual learning scenarios can be ambiguous due to the presence of several words and possible meanings. One possible way to overcome ambiguity is called cross-situational learning (XSL), where information is gathered over several learning trials. Experimental studies of human XSL have shown that cognitive constraints, such as attention and memory limitations, decrease human performance when compared to computer models that can store all available information. In this paper, we approach modeling of human performance with a novel computational XSL algorithm, FAMM (Familiarity preference, Associative learning, Mutual exclusivity, Memory decay), equipped with the four main components motivated by experimental research. The model is evaluated with respect to a number of earlier XSL experiments that probe different aspects of learning. The results show that the model provides a better fit to the behavioral data than the earlier proposed model of Kachergis et al. (2012).

Keywords: cross-situational learning, mutual exclusivity, memory, computational model, familiarity preference

Introduction

Human infants learn their native language in complex interaction within their environment. One important part of language learning is vocabulary acquisition, i.e. learning words and their meanings as they occur in speech spoken by others. This process contains several difficulties that have to be overcome by infants in order to learn to recognize acoustic patterns for words and their corresponding real-world meanings. In this work, we focus on the acquisition of word meanings, assuming that the learner has already solved the word segmentation and lexical decoding problems.

The prevalent view is that infants learn meanings of words by associating acoustic representations of words to some real-life objects or entities ("referents"). A good opportunity for forming such associations is during interaction with an adult, e.g., an adult reading a picture book, where some words in the adult's speech relate directly to some objects or pictures in the shared attentional space between the learner and the adult. However, learning scenarios like this are usually ambiguous, consisting of several words and meanings, where possibly only a few of all possible combinations are correct word-referent pairs.

One way to overcome ambiguity in learning situations is to accumulate evidence over several individually ambiguous situations. This approach is often referred to as cross-situational learning (XSL) (Pinker, 1989). Yu and Smith (2007) have shown that adults use cross-situational statistics to learn word-referent mappings across individually ambiguous learning trials. The findings have also been confirmed for infants (Smith and Yu, 2008) and for children (Suanda, Mugwanya & Namy, 2014).

In laboratory settings it has been shown that human XSL performance is notably below ideal observer performance (see e.g., Yu & Smith, 2012). These limitations may be due to attentional, memory, or other cognitive constraints and a number of behavioral studies paired with computational modeling have been used to investigate the more subtle aspects of the XSL learning process (Yu & Smith, 2012; Kachergis, Yu & Shiffrin, 2012). Although the earlier work has been successful in modeling the behavioral data, parameters of the existing models have been separately fit to each experimental setting (Kachergis et al., 2012), leaving some room for ambiguity concerning whether the data could also be explained with other learning strategies, modeled using similar degrees of freedom.

In this paper, we focus on modeling human XSL performance across a number of tasks involving variation in the number of concurrent tokens, number of repetitions per token, and also variation in whether the spoken words map to one or more visual referents. The first task modeled, that of Yu & Smith (2007) (from here on YU07), alters the difficulty of the learning by varying trial ambiguity and the number of repetitions of words and objects. The other experiment modeled, that of Yurovsky, Yu & Smith (2013) (YUR13), shows a detailed change in the learning results by simple reordering of learning trials, that is presumably caused by more detailed attentive or competitive processes in human learners, whose effects may not be observed in more general learning tasks. Importantly, we present a novel computational model of XSL that attempts to explain all these experimental conditions using a fixed set of model parameters without task-specific fine-tuning.

The present model is compared against an earlier model proposed by Kachergis et al. (2012). The results show that our novel model is capable of accounting for the previously observed effects across fourteen behavioral test conditions, including the detailed ordering effect of Yurovsky et al. (2013), without requiring separate parameter optimization for the different experimental conditions.

Findings from human experiments of XSL

In experimental studies of XSL, human participants are usually presented with a sequence of learning trials, each trial typically consisting of two to four visual objects and the same number of spoken words. The goal of the learner is to acquire correct word-referent mappings during the training trials. The ambiguity across learning trials can be varied in several ways. Each of the displayed referents may correspond to one of the simultaneously presented words, or there may be words that do not have a correct referent. Some studies may include homonymous words (words with several meanings) or synonymous words (several words for one meaning). In addition, the difficulty can be varied by varying the total number of word-object pairs or the number of words and objects present within each trial.

Several learning mechanisms may be present in human XSL and explain the findings in related experiments. Yu and Smith (2012) showed that two competing, seemingly disparate, theories of XSL (hypothesis testing and associative learning) can both replicate findings in human XSL experiments if parameters related to cognitive processes such as information selection and decision strategies are adjusted correspondingly.

Mutual exclusivity (ME) refers to a learner's bias to learn one-to-one mappings between words and meanings. For example, if a learner is presented with a novel word and a novel and a familiar object (the word associated with the familiar object has been already learned), the learner tends to associate the novel word to the novel object (e.g. Markman & Wachtel, 1988). At least 15 to 17 month old infants seem to have developed a bias for ME (Markman, Wasow & Hansen, 2003). However, as humans are able to learn synonyms and homonyms, the strict ME constraint can be violated (e.g. Clark, 1987; Nelson 1988). The gradual violation of the ME rule can also be clearly seen in the experiments of Kachergis et al. (2012).

In XSL, it is crucial that learners can retrieve information from past occurrences of words and referents. Learners' memory can place limitations on how much information from past trials can be remembered. Vlach & Johnson (2013) have studied infants' memory constraints in an XSL task. The findings indicate that 16-month-old infants learn word-referent mappings better if the words and referents are presented in immediate succession (massed) than when they are distributed across time (interleaved). 20-month-olds learn mappings equally well in both conditions, suggesting that the older infants may have had more memory capability to retrieve information over interleaved trials. In this work, we implement a memory constraint that may also explain the findings of several XSL experiments performed by adults.

Existing computational models of XSL

Although it is straightforward to implement XSL if one has unlimited accuracy and memory capacity, the challenge of building a cognitively plausible model is to implement the limitations of human learning correctly. Several computational XSL models have been previously introduced. For example probabilistic models by Frank, Goodman and Tenenbaum (2007) and Fazly, Alishahi and Stevenson (2010) can infer word-to-referent mappings using XSL and can reproduce some general phenomena of human learning, such as fast mapping, but direct comparison to experimental data is not extensively performed.

A recently proposed computational model by Kachergis, Yu & Shiffrin (2012), has shown good fit to data acquired from human experiments in a number of different XSL tasks (Kachergis et al. 2012; Kachergis, Yu & Shifftin, 2013). In their model, attention for familiar word-object pairings competes with attention for uncertain pairings that occur for example when novel words and objects appear. With correct adjustment of the three parameters (balance between familiarity and novelty, total amount of attention and a forgetting factor) to each task, the algorithm matches well with the experimental data (Kachergis et al., 2012). However, the fit over a number of experimental conditions without changing the parameters and thereby the model behavior has not been systematically investigated.

In the novel computational XSL model presented in this paper, we combine a familiarity preference, associative learning, mutual exclusivity and memory decay into one compact model. The model's parameters are optimized to fit the experimental results of YU07 and YUR13 and the model behavior is compared against the Kachergis et al. (2012) model. In addition to analyzing the models' capabilities to explain experiment-specific findings, their fits to the overall pattern of results across all experiments while using a fixed set of parameters are investigated.

The two modeled experiments

Here the two experimental setups whose results we aim to model are described. In both YU07 and YUR13 adult participants faced a task where they were presented with pictures of uncommon objects and heard a sequence of synthetically generated pseudowords. The participants were asked to learn which words were associated with which pictures across the trials.

In **YU07** the conditions of the experiments were varied in terms of number of concurrent words and referents, the overall number of unique words and referents, and the number of repetitions per each word-referent pair. In each trial, a word and its correct referent were always shown together with additional such pairs. The five experimental conditions consisted of the following: E1) 2 words and 2 referents shown concurrently from a set of 18 unique words/referents, each pair occurring a total of six times (2x2 / 18 words / 6 repetitions), E2) 3x3 / 18 / 6, E3) 4x4 / 18 / 6, E4) 4x4 / 9 / 8 and E5) 4x4 / 9 / 12. In the test phase, participants were presented with one word and four pictures familiar from the training phase and were asked to point out the correct referent among these.

YUR13 training. Here we investigate the first three experimental conditions of Yurovsky et al. (2013) (E6, E7 and E8 from here on) as the fourth one simply used a

different methodology to probe human performance trial-by-trial in the third condition. All their experiments consisted of four objects and words per trial, but not all object-word pairs had a correct one-to-one correspondence. More specifically, there were single words with only one correct referent, double words with two correct referents (i.e. homonyms), and noise words with no associated referents. Single words occurred six times with their correct referents and double words occurred six times with each of their correct referents. Each experiment consisted of 27 trials in total.

In E6 single words always co-occurred with their correct referent, double words always co-occurred with both of their referents, and noise words were randomly used to complete the four-word sequence when double words were present.

In E7 there were no noise words and within each trial, single words always co-occurred with their correct referent, and double words always co-occurred with only one of their correct referents. The ordering of trials was randomized - at every trial it was equally likely to have a double word occurring with its first or second referent. E8 was equal to E7 except that the ordering of the trials was changed so that double words always occurred with their first referent on the first half of the trials (early referents) and with their second referent on the second half of the trials (late referents).

YUR13 testing, participants were tested for their knowledge of correct pairings by letting them hear each of the single and double words, and rank four presented objects in order of likelihood of being the correct referent for the heard word. Importantly, both correct referents were present for double words and the only correct referent for single words, in addition to foil referents. The participants were considered to know the correct referent for a single word if the correct referent was ranked the first (*single* condition). They were considered to know one of the two referents for a double word if either of them was ranked the first (*either* condition), and both of the two referents if both of the correct referents were ranked in positions one or two (*both* condition).

In E6 and E7 Yurovsky et al. (2013) found that participants were significantly less likely to learn both referents of double words than the only referent of single words, indicating that the two referents of double words seem to inhibit each other somehow across trials through global competition. Surprisingly, the learning of double words was greatly enhanced in E8 due to simple ordering of the stimuli and the participants were no longer more likely to learn one referent of a single word than both referents of a double word. Yurovsky et al. offered an explanation that as global competition should inhibit the learning of late referents, perhaps the smaller ambiguity during the first half of training leads to strong learning of single word referents, and competition within a trial (*local competition*) supports late referent learning as the already learned pairings can be excluded from the set of potential new associations. In this paper we offer an alternative explanation for the finding, supported by our computational learning model.

The new FAMM model of XSL

In order to build an XSL model that would fit experimental data without the need to optimize parameters separately to each individual learning task, we have investigated a compact model consisting of hypothesized learning components that might explain the findings of YUR13 (but also YU07). The present model is constructed taking into account the most important findings from human experiments as explained above. The four main components of the FAMM model are:

***F**amiliarity preference.* We hypothesize that in XLS learning there exists a strong familiarity preference towards already seen word-meaning associations. If the learner remembers that a word and an object have co-occurred in a previous trial, and they co-occur again, the learner substantially strengthens the association between the two.

***A**ssociative learning.* The associative learning component associates every object to every word within a trial with a relatively small weight. This sort of component is needed to loosen the ME rule in order to learn homonymous or synonymous mappings within a trial (see introduction above).

***M**utual exclusivity.* Based on experimental evidence (see introduction), the learning model should include an ME component so that the learner has a bias to create one-to-one mappings between words and their meanings. In FAMM, this bias works within each trial so that if any word (referent) within a trial is already associated with any referent (word), no further associations are made except for a small random value induced by the associative learning component

***M**emory decay.* We hypothesize that the learner cannot remember the previous training trials perfectly and that detailed information on the trials is lost rather rapidly. Intuitively, it would seem that on the second trial, if the learners hear a word or see an object that were present also on the first trial, they seem familiar, and they thus certainly have co-occurred previously, because only one training trial has been seen this far. When more trials appear, remembering if seen objects or heard words have co-occurred *within* any previous trial should become difficult, even though individual words or objects might seem familiar. The effect of memory decay is also supported by the findings of Vlach & Johnson (2013), where memory capacity has been offered as an explanation for the difference in performance between 16 and 20 month olds in cross-situational word-object learning task.

Model implementation

The task of the learner is to learn an association matrix **A** between the referents and the words presented during the trials. The rows of **A** correspond to the seen visual objects (=referents) and the columns to the heard words. **A** is initialized with zeros (no associations), and it is updated on every trial based on the heard words, seen objects, and the previous values of **A**.

Associations between the observed words and objects are generally strengthened by a constant value depending on a parameter α. Mutual exclusivity can however inhibit the strengthening of certain pairs. A small random component depending on a parameter β is also added to all pairs to account for associative learning. Non-linear memory decay depends on the third parameter γ.

Training. On current trial t, where a total of N_O objects $\{o_1, o_2, \ldots, o_{No}\}$ and N_W words $\{w_1, w_2, \ldots, w_{Nw}\}$ are present, their association scores are updated to matrix **A** as follows

$$\mathbf{A}^{t}_{o_i,w_j} = D \cdot \left(\mathbf{A}^{t-1}_{o_i,w_j} + F + M + R\right), \quad (1)$$

where D, F, M and R refer to memory decay, familiarity, mutual exclusivity and random components for each association correspondingly. The familiarity component adds a value of 1 to the association between o_i and w_j if it is remembered that they have co-occurred on any previous trial:

$$F = \begin{cases} 1, & \text{if } \mathbf{A}^{t-1}_{o_i,w_j} > 0 \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

The mutual exclusivity component M makes sure that the objects and words in the current trial that are remembered to have been associated in some previous trial, will not be associated to any other words or objects within the current trial. They are thus considered to be a confirmed pair.

$$M = \begin{cases} 0, & \text{if } \sum_{w=w_1}^{w_{Nw}} \mathbf{A}^{t-1}_{o_i,w} > 0 \text{ or } \sum_{o=o_1}^{o_{No}} \mathbf{A}^{t-1}_{o,w_j} > 0 \\ \alpha / N_c, & \text{otherwise} \end{cases} \quad (3)$$

where N_c refers to the number of possible combinations of word-object pairs in a trial, leading to the assumption that when a trial has less words and objects, the learner is able to pay more attention to the possible combinations.

The random component $R \sim U(-\beta/2, \beta/2)$ assigns a random association value between every object and every referent. This models the noise present in learning situations, attention and brain processes. It also helps to bring variety to the forgetting process. Because of the non-varying memory decay component, without any kind of noise, all object-word pairs in the memory with equal association values would be forgotten equally fast.

After the update phase, the non-linear memory decay factor takes place before the next trial is presented, and is implemented using a formula

$$D = \left(\tanh\left(\gamma \cdot \left(20 \cdot \left(\mathbf{A}_{t-1}(o_i, w_j) + F + M + R\right) - 10\right)\right) + 1\right) / 2.01 \quad (4)$$

The function based on hyperbolic tangent expresses the decay factor with which the current association values are multiplied. The decay factor depends on the association values after the update phase, so that strong associations decay less rapidly than weak associations. Figure 1 shows the updated association values as a function of the old association values (values before the decay function) when γ is 0.1, 0.3 and 0.5. Note that the division by 2.01 in equation (4) makes sure that even large association values are decayed minimally – the largest decay factor for association values over one is 0.995. In the end of each trial, values of **A** below 0.01 are set to zero to model complete forgetting.

Testing. When tested for word-referent associations in the YU07 experiments, FAMM always selects the most strongly activated referent, given the test word..In YUR13, the model ranks the associated referents as in the original experiment. The choices for each test trial are chosen equally to the original experiments, but in YUR13 we test each single and double word six times to account for variability caused by the randomization of the foils in each test trial. After all training trials, before the testing phase, a small amount of noise (U(0, 0.01)) is added to **A** in order to randomize the selected referent for the test word in case several cells of **A** have the same value.

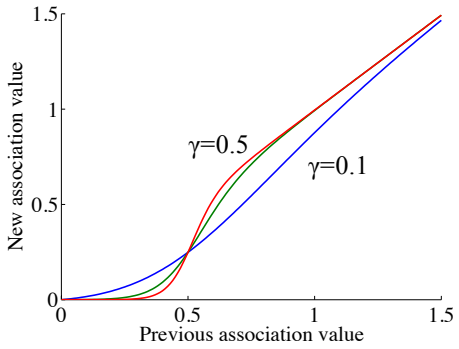

Figure 1. The effect of the memory decay function. x-axis = the original value in association matrix **A**. y-axis = the value after memory decay. Values γ = 0.1, 0.3 and 0.5 are shown.

Experiments

Experimental setup

In this paper, computational modeling of the experimental setups of YU07 and YUR13 is presented. The reason for choosing these two studies is that the former investigates the basic properties of XSL under varying degrees of concurrent words, referents and the overall number of trials while the latter study reveals interesting details about competition between associations during learning.

Modeling details

The performance of the new XSL model is compared to the performance of the XSL model by Kachergis et al. (2012). Both of the models are fitted to the experimental data of Yu & Smith (2007) and Yurovsky et al. (2013) by performing a grid-search for optimal parameter values. Both models have three hyperparameters that have notable impact on the model performance and therefore the search is performed across the relevant range of all these parameters.

In the testing phases of the experiments, Kachergis et al.'s algorithm makes hypotheses for words' referents proportionally to their association values (Luce's choice rule) as described in Kachergis et al. (2012).

The optimization is performed at two levels: 1) individually for each of the eight simulated behavioral

experiment conditions (five conditions from YU07, and three conditions from YUR13), and 2) at a global level by finding a single set of optimal parameters across all fourteen data points (note that each condition in YUR13 has results for three different token types, yielding 3 x 3 = 9 data points). In both cases, the optimization criterion is the RMSE between the average model output and the means of the reported behavioral data.

The search grid we used for the Kachergis et al. (2012) algorithm (see parameter descriptions in the original paper) was α = [0.7, 1] (step-size = 0.02), χ = [0.1, 8] (ss = 0.2) and λ = [0.1, 8] (ss = 0.2). For FAMM α = [1, 3.4] (ss = 0.2), β = [0.1, 0.36] (ss = 0.02) and γ = [0.1, 0.5] (ss = 0.05). With each set of parameter values, the experiments were run 15 times, and the RMSE between the averaged results and the experimental data point means was stored. After the optimization, accounting for the variance over several runs of the algorithm and in order to get a comparable result to the experimental data, the results are averaged across P runs, where P represents the average number of participants in the eight original experimental conditions of YU07 and YUR13 (P = 38). The RMSE between the 14 data points of the experimental data average and the model average is calculated after every run. This comparison is repeated 20 times in order to obtain means and deviations of the models' performances.

Results

As a result of optimizing across all eight experiments, the optimal hyperparameter values for Kachergis et al. algorithm were α = 1, χ = 0.1 and λ = 6.5, and for the FAMM model α = 2.60, β = 0.16 and γ = 0.25. The experiment specific and global fits between the two models and all eight experimental conditions (14 data points) are shown in Table 1 and the means and standard errors of all 760 runs pooled in Figure 2.

The overall finding is that both models fit nearly perfectly to almost all individual experiments when the parameters are optimized specifically for each condition. However, differences are seen when the parameters are not allowed to change between the tasks. More specifically, FAMM leads to a significantly better overall fit (Wilcoxon rank-sum test, W = 210, $p << 0.001$) with almost half of the global error of the Kachergis et al. (2012) model.

In addition to that, the FAMM is capable of following the pattern in double word learning across E7 and E8 with notably enhanced learning of both meanings of double words in E8 (over 760 runs, E7: *Mdn* = 25.00, E8: *Mdn* = 36.11, W = 480000, $p << 0.001$). In contrast, double word learning for Kachergis et al. model is significantly better in E7 than in E8 (E7: *Mdn* = 30.56, E8: *Mdn* = 27.78, W = 599000, p = 0.013).

If the hyperbolic tangent decay function of FAMM is replaced with a simple linear decay factor $D = \gamma$, where γ is optimized in range [0.1, 1], a global RMSE error of 64.77 is achieved (when γ = 0.2, i.e. fast memory decay). Double-words are not learned better in E8 than in E7 anymore (1000 runs, E7: *Mdn* = 22.22, E8: *Mdn* = 22.22, W = 1006741, p = 0.63). The familiarity and hypothesis testing components thus do not seem to suffice to explain the increase in learning both referents of double words in E8. Replicating the effect seems to require stronger memory decay for weak associations. We hypothesize that the early referents of double words are learned better in E8 because their co-occurrences are packed closer together on the first half of the trials making it more likely that an early referent pair is remembered on its new occurrence, when the familiarity principle strongly stores the association into memory.

Table 1. Errors and correlations between the two models and the experimental data.

	Measure	Global fit (SD)	Experiment-specific fit
FAMM	RMSE	**22.81 (2.75)**	**4.30**
	Lin. corr. r	0.96 (0.01)	1.00
	Rank corr. ρ	0.96 (0.02)	1.00
Kachergis et al. (2012)	RMSE	41.24 (0.73)	5.68
	Lin. corr. r	0.83 (0.01)	1.00
	Rank corr. ρ	0.74 (0.03)	1.00

Conclusions

Existing computational models of cross-situational learning have generally replicated some general patterns in human learning without extensive comparison to experimental data (Fazly et al., 2010; Frank et al., 2007), or their parameters have been adjusted to individual experimental conditions with a risk of overfitting to data (Kachergis et al. 2012; Kachergis, Yu & Shifftin, 2013). We have investigated what learning components should be included in a computational model of XSL in order to match experimental data more globally, i.e. optimizing one set of parameters to a larger amount of experimental data.

We presented a novel computational XSL algorithm, FAMM, that can replicate experimental results of Smith &

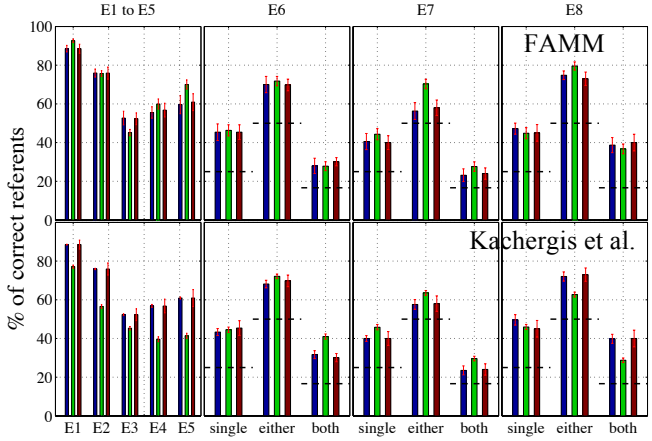

Figure 2. Modeled experimental data for all experimental conditions. The blue left-most bars indicate the model fit with parameters optimized for each experimental condition individually. The green middle bar indicates the model fit with one set of globally optimal parameters. The red bar on the right shows the results of YUR13. Standard error bars are shown with red, and chance level performance with horizontal lines.

Yu (2007) and Yurovsky et al. (2013) better than a comparable model (Kachergis et al., 2012). Essential components of the algorithm are a nonlinearly decaying memory trace of associations, a strong familiarity preference, mutual exclusivity component and a small random association component for all word-referent pairs. The strong familiarity preference "confirms" remembered associations and the nonlinear memory decay makes sure that strong associations, formed with the help of the strong familiarity preference, stay in the memory longer than weak associations that are more likely to be made with chance and are thus more likely to be incorrect.

Special attention was paid to a finding of Yurovsky et al. (2013), where both referents of double words were learned more reliably when the first referent was presented only during the first half of the training trials, and the second referent only during the second half (see the experimental setup section). Only the FAMM algorithm was able to replicate the finding. Yurovsky et al. (2013) offered a possible explanation that mutual exclusivity may lead to the difference between the two conditions. This is because the learner has learned single words more strongly due to reduced ambiguity during the first half of the experiment, and can therefore exclude single words during the second half and pay more attention to the new double words and their referents. The present simulations indicate that incorporating mutual exclusivity alone may not be enough to replicate the experimental findings. Instead, the study with FAMM suggests that the relatively large boost observed in learning of double words in E8 is caused by the following mechanism: Since the frequency of double words co-occurring with their first referent is about twice as big in E7 than E8 (experiments 2 and 3 in the original paper) during the first half of the learning trials, the early word-referent pairs are more likely to get "consolidated" due to the strong familiarity preference before their co-occurrences during the previous trials become forgotten. The learner simply remembers more of the early pairs during the second half of learning. In contrast, the memory decay has more severe consequences in the interleaved conditions (E6-7).

The most important component of the FAMM model when compared to the Kachergis et al. model seems to be the nonlinear memory decay component that leads to the detailed finding of Yurovsky et al. (2013) considering E8. Without this component the model does not reach Kachergis et al. model's performance as is shown in the results section. The effect of the memory component also suggests that participants in XSL experiments forget seen associations rapidly, and in order to remember certain word-referent pairs, they should be repeated in nearby trials. Also in FAMM, the strength of the association update on each trial depends on the number of possible combinations between the presented words and referents (see eq. (3)), making the model more flexible towards different XSL conditions, whereas in Kachergis et al. model the parameters are independent of the trial difficulty.

Acknowledgments

This research was funded by the ETA Graduate School of Aalto University, Finland, the ERC starting grant project ABACUS, grant number 283435 and the Academy of Finland. The authors would like to thank Bart de Boer, Hannah Little and Bill Thompson for insightful comments.

References

Clark, E. (1987). The principle of contrast: A constraint on language acquisition. *Mechanisms of language acquisition*, 1–33.

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

Frank, M. C., Goodman, N. D., & Tenenbaum, J. B. (2007). A Bayesian framework for cross-situational word-learning. In J.C. Platt, D. Koller, Y. Singer, & S. Roweis (Eds.), *Advances in Neural Information Processing Systems, Volume 20* (pp. 1212–1222). Cambridge: MA: MIT Press.

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

Kachergis, G., Yu, C., & Shiffrin, R. M. (2013). Actively learning object names over ambiguous situations. *TopiCS*, *5*, 200–213.

Markman, E. M., & Wachtel, G. F. (1988). Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive psychology*, *20*(2), 121–157.

Markman, E. M., Wasow, J. L., & Hansen, M. B. (2003). Use of the mutual exclusivity assumption by young word learners. *Cognitive psychology*, *47*(3), 241–275.

Nelson, K. (1988). Constraints on word learning? *Cognitive Development*, *3*(3), 221–246.

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

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

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

Vlach, H. A., & Johnson, S. P. (2013). Memory constraints on infants' cross-situational statistical learning. *Cognition*, *127*(3), 375–382.

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. (2012). Modeling cross-situational word-referent learning: Prior questions. *Psychological Review*, 119, 21–39.

Yurovsky, D., Yu, C., & Smith, L. B. (2013). Competitive processes in cross-situational word learning. *Cognitive Science*, 37, 891–921.