Modeling individual performance in cross-situational word learning

Yung Han Khoe (y.khoe@student.ru.nl)
Centre for Language Studies
Radboud University Nijmegen

Amy Perfors (amy.perfors@unimelb.edu.au)
School of Psychological Sciences
University of Melbourne

Andrew T. Hendrickson (a.hendrickson@tilburguniversity.edu)
Department of Cognitive Science & Artificial Intelligence
Tilburg University

Abstract

What mechanisms underlie people's ability to use cross-situational statistics to learn the meanings of words? Here we present a large-scale evaluation of two major models of cross-situational learning: associative (Kachergis, Yu, & Shiffrin, 2012a) and hypothesis testing (Trueswell, Medina, Hafri, & Gleitman, 2013). We fit each model individually to over 1500 participants across seven experiments with a wide range of conditions. We find that the associative model better captures the full range of individual differences and conditions when learning is cross-situational, although the hypothesis testing approach outperforms it when there is no referential ambiguity during training.

Keywords: Cross-situational word learning; language acquisition; Zipfian distributions

Introduction

The ability to acquire language is not only a fundamental part of what makes us human, but a mystery: how do we accomplish it given the complexity of the learning task? Even in an apparently simple task like word learning, many real-world contexts involve multiple possible referents for any one label (Pinker, 1984). How can a learner figure out which referent corresponds to which label? One suggestion is that people can leverage the statistics that come from observing multiple ambiguous presentations of words and objects. This sort of cross-situational word learning has been demonstrated in both children and adults (Yu & Smith, 2007; L. Smith & Yu, 2008). However, there is still considerable debate about what mechanisms underlie cross-situational word learning and what representations are learned (Kachergis & Yu, 2018).

One major theory of cross-situational learning, known as the associative framework, proposes that people track detailed word-object co-occurrence statistics across many presentations (Vouloumanos, 2008; Yu & Smith, 2007). By contrast, the hypothesis testing framework suggests that people track at most one word-object pair theory for each word (or object) and update these hypotheses during learning (Medina, Snedeker, Trueswell, & Gleitman, 2011). A number of computational models have been developed based on both frameworks but no consensus has emerged about which account better describes people's learning. We argue that this has occurred, at least in part, because of a focus on modeling aggregate rather than individual data, and because existing experiments have not varied the range and variety of learning conditions sufficiently to differentiate the models.

Here we present and analyze data from seven different experiments with over 1500 participants that vary on a number of factors including vocabulary size, level of ambiguity, length of training, distributional structure, and task. We fit each person's data to both the associative and hypothesis testing models described in the following sections. Our results suggest that associative accounts provide the best fit in almost all cases, unless there is no ambiguity during learning and the learning is thus no longer strictly cross-situational.

Associative framework

Associative models propose that people learn word meanings by tracking the frequency with which words and objects co-occur across multiple ambiguous presentations. The representation is a large word-object matrix in which each cell contains the associative strength between one word and one object (Vouloumanos, 2008; Yu & Smith, 2007). This basic framework has been applied widely, and the model we implement here is one of the most widely used (Kachergis et al., 2012a). It provides a compelling account of human behavior across studies that vary the number of late repetitions (Kachergis et al., 2012a) and if learning is passive or active (Kachergis, Yu, & Shiffrin, 2012b; Kachergis & Yu, 2018).

Formally, the goal of the model is to update the association strength between a word (w) and object (o) within an association matrix ($M_{w,o}$). It incorporates several psychologically-motivated parameters that specify the total amount of updating on each trial (χ), memory fidelity (α), and the bias towards updating the association strength of uncertain versus already familiar words and objects (λ) with uncertainly of an item quantified as the entropy across all association strengths for that item.

Hypothesis testing framework

As an alternative to the memory-intensive associative framework, Medina et al. (2011) outlined a more minimalistic approach based on storing only a single hypothesis for each word. The hypothesis represents a guess about the referent of the word, and is replaced if it is inconsistent with new training trials or fails to be recalled when the word is present.

Vocabulary	Ambiguity	Guessing	Presentations	Distribution	Length Relationship	N	Source
12	3	Yes	108	Uniform	Only one syllable	48	H&P (2018) Exp 1
12	3	Yes	108	Zipfian	Only one syllable	72	H&P (2018) Exp 1
32	4	Yes	244	Uniform	Uniform	79	H&P (2018) Exp 2
32	4	Yes	244	Zipfian	Correlated	81	H&P (2018) Exp 2
32	4	Yes	244	Zipfian	Random	80	H&P (2018) Exp 2
32	1	NA	244	Uniform	Uniform	74	H&P (2018) Exp 3
32	1	NA	244	Zipfian	Correlated	77	H&P (2018) Exp 3
32	1	NA	244	Zipfian	Random	86	H&P (2018) Exp 3
28	4	Yes	240	Uniform	Uniform	171	Exp 1
28	4	Yes	240	Zipfian	Random	166	Exp 1
40	4	Yes	240	Uniform	Uniform	71	Exp 2
40	4	Yes	240	Zipfian	Random	90	Exp 2
28	4	No	240	Uniform	Uniform	82	Exp 3
28	4	No	240	Zipfian	Correlated	84	Exp 3
28	1	NA	240	Uniform	Uniform	159	Exp 4
28	1	NA	240	Zipfian	Correlated	151	Exp 4

Table 1: **Overview of experimental structure**. This table describes all of the experiments whose data we fit. *Vocabulary* indicates the number of unique word-object pairs to be learned (which also corresponds to the number of objects present during the test phase). *Ambiguity* indicates the number of objects present on each training screen. *Guessing* indicates whether learning was passive (just watching) or active (if participants were required to submit a guess after each word during training). *Presentations* indicates the total number of training trials. *Distribution* indicates the frequency distribution of the words and objects across the experiment. *Length Relationship* indicates the relationship between the length of words and their frequency during training, with more frequent words being shorter in the *Correlated condition*. *N* indicates number of complete participants. *Source* indicates the source of the data set: H&P (2018) denotes Hendrickson and Perfors (2018).

We evaluate the Propose-but-Verify hypothesis testing model (Trueswell et al., 2013), a popular extension of the original Medina et al. (2011) formulation, which captures children's word learning behavior well (Woodard, Gleitman, & Trueswell, 2016; Aravind et al., 2018). The model involves a two-stage process. Upon initially being exposed to a word, the model chooses an object from the as-yet-unmapped objects in that trial and maps it to that word to form a word-object hypothesis. The initial probability of later recalling that mapping is denoted by a free parameter $\alpha_{initial}$. On each subsequent exposure to the word, if the model recalls the hypothesis and the corresponding object is present, the probability is updated to a different memory strength indicated by another free parameter, $\alpha_{confirmed}$. If the hypothesis fails to be recalled or the corresponding object is not present, a new hypothesis is established with an unmapped object.

Model comparisons

Many previous papers have compared these two modeling approaches in terms of how well they fit experimental data (e.g., K. Smith, Smith, & Blythe, 2009; Kachergis et al., 2012b; Rasilo & Räsänen, 2015; Kachergis & Yu, 2018; Aussems & Vogt, 2018; Stevens, Gleitman, Trueswell, & Yang, 2017). Despite this effort, no consensus has emerged. One reason may be the focus on modeling aggregate performance using one optimal set of parameter values per model for all learners, which ignores individual differences. This approach may favor highly stochastic models that can fit different people's responses with a single parameter, rather than models that can fit the behavior of more people using individual parameter values. Moreover, comparison studies commonly fit these models to experiments that involve relatively few learners, and have a small number of conditions which do not capture the variation across conditions in the literature. Finally, such studies tend to use uniform word frequencies that do not reflect the highly-skewed distribution of words in natural language, which limits the generalizability to real-world word learning (Hendrickson & Perfors, 2018).

In this paper we address these issues by evaluating a hypothesis testing model and an associative learning model against experimental data involving over 1500 participants and spanning the broad range of conditions shown in Table 1. We varied the distribution of the words and objects, the size of the vocabulary to be learned, whether the task was passive or active, the number of presentations during training, the level of ambiguity during learning, and the relationship between the length of the word and word frequency. We fit parameter values for each learner by optimizing the log-likelihood of model response probability for each of the word-object test trials. When comparing models, we penalize for additional parameters by converting the log likelihood to AIC values (Akaike, 1974).

Experiments

The empirical data that we use for model evaluation includes data from the eight conditions from Hendrickson and Perfors (2018) in addition to eight additional new conditions. We describe each in turn.

Hendrickson and Perfors (2018)

The goal of the work in Hendrickson and Perfors (2018) was to explore cross-situational learning when the words followed either a ZIPFIAN or a UNIFORM distribution. The first experiment involved presenting participants with a small vocabulary of words in one of the two distributions. The second increased the vocabulary and ambiguity level, while adding a condition in which the length of the word was negatively correlated with word frequency (shorter words were more frequent). The third evaluated the effect of removing ambiguity during training.

Procedure. Each experiment consisted of a training phase and a test phase, though Experiment 1 repeated these phases multiple times. During training, participants viewed either 3, 4, or 1 objects on the screen at once while they heard the words for each object presented one at a time in random order. In experiments with ambiguity during training, participants were asked to guess which object each word matched. At test, people were shown all of the items at once and asked to select the matching object for each word. They were not given feedback during training or test.

Conditions. In the UNIFORM conditions the words and objects all occurred with the same frequency, while in the ZIPFIAN conditions a few words and objects occurred very frequently and many words occurred very infrequently or only once across training. The pairing of words and objects and trial order was randomized across participants.

Materials. Words varied in length from one to three syllables and were designed to sound English-like as well as be maximally distinct from each other. They were generated by the AT&T Natural Voices Text-to-Speech tool (Crystal voice). The objects were selected from the NOUNS image corpus (Horst & Hout, 2015) and each image was 150x150 pixels displayed against a white background. Hendrickson and Perfors (2018) contains the full set of stimuli.

Participants. Their 597 participants were recruited from Amazon Mechanical Turk (AMT). Our four additional experiments (with 974 participants) were also run on AMT, paying US$3.25 for the $\sim$20 minute task.

Experiment 1

Experiment 1 provides a near replication of Experiment 2 of Hendrickson and Perfors (2018), a design aimed to approximate learning conditions when the meaning of words is ambiguous. There were two minor differences. First, their experiment included four single-presentation items in order to check for participant cheating; we omitted those in order to ensure that the UNIFORM distribution contained no low-frequency items. As a result, we had 240 rather than 244 total presentations and 28 rather than 32 test items. Second, we did not include their second ZIPFIAN condition, in which the length of the word and word frequency was correlated. 337 individuals provided complete data, half in the UNIFORM condition and half in the ZIPFIAN condition.

Experiment 2

A number of simulation studies have suggested that increasing the number of items to be learned should be particularly challenging for learners in Zipfian environments (Vogt, 2012; Reisenauer, Smith, & Blythe, 2013). In Experiment 2 we therefore replicated the design of Experiment 1 but with 40 unique word-object pairs instead of 28. We presented each word slightly less frequently in order to match the total number of word-object presentations in Experiment 1. In the test phase 40 rather than 28 objects were displayed, resulting in a more difficult test. Complete data was collected from 161 individuals, with roughly half assigned randomly to the UNIFORM and ZIPFIAN conditions.

Experiment 3

In all of the experiments so far, participants have been required to respond by selecting a best-guess object after each word was presented during training. However, recent work has suggested that forcing people to guess may influence the representation that they learn (Aussems & Vogt, 2018). We address this possibility in Experiment 3, which is identical to Experiment 1 but removes the obligation to guess during training. Instead of waiting for a guess after each word, the next word is played automatically after 2000 ms.

The other difference from Experiment 1 is that the length of each word was correlated with its frequency in the ZIPFIAN condition, as is found in natural language and Hendrickson and Perfors (2018). Complete data was collected from 166 individuals, with roughly half assigned randomly to the UNIFORM and ZIPFIAN conditions.

Experiment 4

Experiment 4 provides a near replication of Experiment 3 of Hendrickson and Perfors (2018), whose goal was to approximate learning when the meaning of words was unambiguous. The only differences, as in Experiment 1, were that we removed the four "cheating check" items and thus had 240 presentations and 28 test items, and we had only one ZIPFIAN condition in which word length was correlated with frequency. Additionally, since there was no ambiguity during training, participants were not required to guess. Instead, the timing of item presentation matched Experiment 3. Complete data was collected from 310 individuals, with roughly half assigned to the UNIFORM and ZIPFIAN conditions.

Model Fitting

Both models were fit to the individual data of each person independently by minimizing the negative log likelihood across all responses in the test phase. Every person participated in exactly one condition and thus parameters were not constrained across conditions in any way. For the associative model, the likelihood of a correct answer was determined for each word by dividing the associative mass on the correct object by the total associative mass across all objects. For

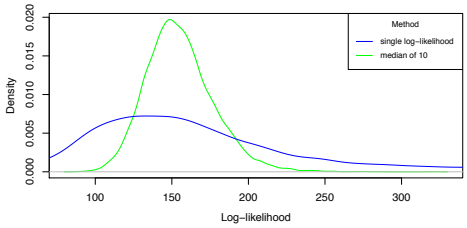

Figure 1: **Distribution of log-likelihood vs. median log-likelihood values for the hypothesis testing model.** The density plot shows the different distributions of 10,000 simulations produced by the hypothesis testing model for one individual participant with a constant set of parameters when using either a single log-likelihood value or the median of 10 such values. Using the median of 10 values results in a substantial increase in stability.

the hypothesis testing model, the likelihood was given by the stored probability value for a correct pairing, smoothing zero probability values to 0.0001.

Interestingly, the models differed widely from each other in the variability of the responses probabilities predicted at test. Given a fixed training trial order and a single set of parameter values, the associative model has no stochastic aspect to how the representation is formed. Therefore, the likelihood of a set of responses given a set of parameters is stable and parameter estimation was straightforward.[1]

In contrast, the hypothesis testing model is decidedly random about which words are paired with objects when forming hypotheses. This results in the production of markedly different representations and thus likelihoods from one simulation to the next, even when the training trials and parameter values are constant across runs. In order to address this issue, we performed ten simulations for each set of parameters during the optimization process and used the median log likelihood across the simulations. This required the use of a particle swarm optimization algorithm to determine the optimal parameters, which is more robust to less smooth optimization problems.[2] It was notable that across the range of 10 likelihood values for a set of parameters, the best value was markedly better than the median likelihood value, suggesting that optimization routines that rely on the best likelihood given a set of parameters can overestimate the expected fit to data of a set of parameters, especially for the hypothesis testing model. In addition, the median likelihood from 10 simulations of the hypothesis testing model produces considerably more stable estimates relative to a single simulation (Figure 1).

The number of parameters differ between the two models, with the associative model containing three (α, χ and λ) and the hypothesis testing model containing only two ($\alpha_{initial}$ and $\alpha_{confirmed}$). We therefore penalized for model complexity by converting the log likelihood scores to AIC values; lower AIC scores indicate a better fit to the data after taking the number of free parameters into account.

[1]The optimal parameter values were derived using the default settings for the optimize function in the SciPy package in Python 3.

[2]Fitting was done using the PSO package in Python 3 using a swarm size of 1,000 and a maximum of 50 iterations.

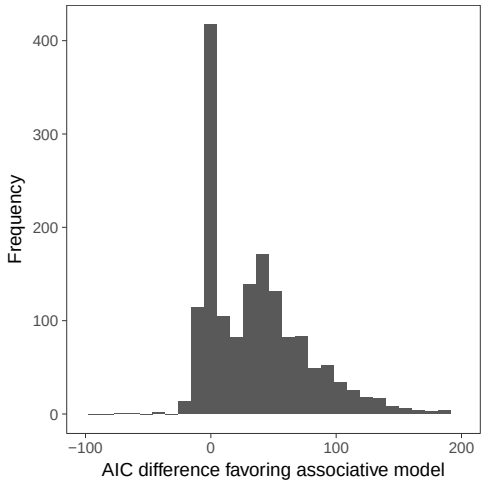

Figure 2: **Overall model performance.** The histogram shows the difference between the AIC score for the hypothesis testing model and the AIC score for the associative model for each person across all datasets. Positive scores (which made up 74% of the data) indicate that the associative model had a lower AIC score than the hypothesis testing model and thus accounted for the data better.

Results

Figure 2 shows the overall performance of the two models across all individuals. The AIC scores of the associative model are lower than the hypothesis testing model for 74% of participants. Since a lower AIC indicates better fit, this suggests that for most people the associative model provides a better account of their performance. Next, we turn to exploring exactly when and where each model does best.

Ambiguity. As the top row of Figure 3 illustrates, the performance of the two models strongly depends on the degree of ambiguity during training. In conditions with any degree of ambiguity during training (by presenting three or four items on each training screen, rather than individually), the associative model is a better fit in virtually all cases (97% of participants). However, the opposite is true when there is no ambiguity: when only one item was shown at a time, the hypothesis testing model was favored for 68% of participants.

Word frequency. The near unanimous advantage for the associative model in ambiguous learning conditions suggests that the only differences in model performance due to word frequency might occur in the conditions without ambiguity. In the conditions with only one item per screen (top left panel of Figure 3), the hypothesis testing model (red points) is highly preferred (96% of the time) when the distribution is UNIFORM. When the word frequency distribution is ZIPFIAN (blue points), the two models are roughly even (52% of participants are better fit by the associative model).

Vocabulary size. The impact of vocabulary size on model performance differs based on the ambiguity of word meaning during training (bottom row, left and center panels in Figure 3). When word meaning is ambiguous during training, the hypothesis testing model does particularly poorly as the vocabulary size increases. (Note that we do not show the 3-item

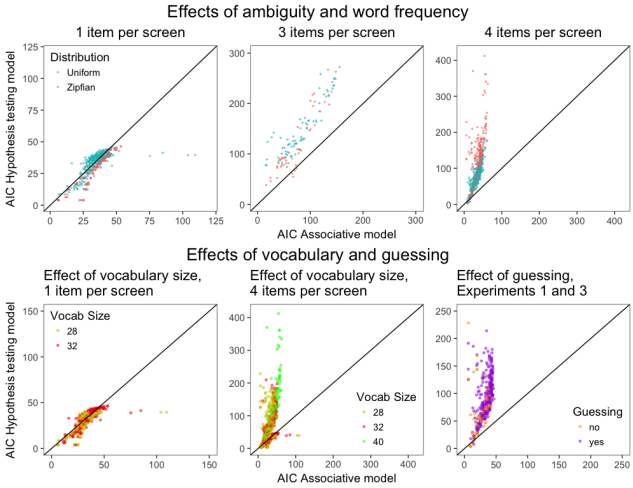

Figure 3: **Model comparisons for different effects.** All panels show the AIC scores for the hypothesis testing model (y axis) plotted against the AIC scores of the associative model (x axis). Values above the identity line indicate that the AIC score of the associative model is better (lower). *Top:* A comparison of the effects of ambiguity and word frequency. Each red dot shows a participant in the UNIFORM condition; blue dots represent people from the ZIPFIAN condition. Each panel shows a different level of ambiguity during training (1, 3, or 4 items on screen at once). The associative model does much better whenever there is ambiguity (and regardless of distribution), while the hypothesis testing model does slightly better when there is only one item on screen during learning. *Bottom left and center:* Evaluation of the effects of vocabulary size, broken down by degree of ambiguity. When training is unambiguous, there is no consistent effect of vocabulary size on performance; when it is ambiguous, the hypothesis training model appears to perform especially poorly when there are more words to be learned. *Bottom right:* Regardless of whether participants were passive or active learners, the AIC favored the associative model.

ambiguous case because vocabulary size in those conditions was constant at 12 items). Within the unambiguous training conditions there does not appear to be any consistent effect of vocabulary size on model performance.

Guessing. In order to evaluate the prediction that forcing participants to guess during training can bias participants to adopt hypothesis testing representations (Aussems & Vogt, 2018), the bottom right panel of Figure 3 shows the performance of both models as a function of whether participants had to guess or not. Here we include only data from Experiments 1 and 3, similar experiments that differ in whether guessing occurs. Both have a vocabulary size of 28 and ambiguous training (although the relationship between word length and word frequency differs between the two ZIPFIAN conditions). Across all conditions the associative model consistently outperforms the hypothesis testing model.

Fitted Parameters

In addition to being useful for model comparison, the best-fitting parameters across all participants (shown in Figure 4) provide us with several deeper insights. First, the distribution of parameter values can tell us something about the distribution of individual differences across the population. For example, the distributions of α and χ for the associative model are highly skewed and show ceiling and floor effects, which suggest these parameters might not capture meaningful variation across individuals. By contrast, the distributions of the memory strength parameters for the hypothesis testing model both display a unimodal peak around relatively high values with a long tail of low values for some participants. This suggests a high level of population variance or reflects the inherent stochasticity of the hypothesis testing model.

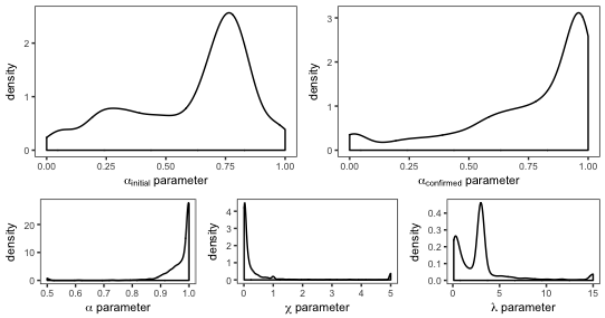

Figure 4: **Distribution of best fitting parameters across all participants**. *Top row*: Hypothesis testing model parameters $\alpha_{initial}$ (initial probability of recalling a mapping) and $\alpha_{confirmed}$ (later probability of recalling a mapping). *Bottom row*: Associative model parameters α (memory decay rate), χ (amount of updating for each trial), and λ (bias towards updating uncertain words and objects).

It is also useful to compare our best-fit parameters to the reported values from other studies, which generally fit aggregate data or use other fitting metrics. The distribution of our fitted values for the two hypothesis testing model parameters are generally higher than those reported by Trueswell et al. (2013) in their two experiments: $\alpha_{initial}$ values of 0.26 and 0.60, and $\alpha_{confirmed}$ values of 0.71 and 0.81. On average, our best-fit parameter values were higher and show less difference between the initial and confirmed memory strength. It remains an open question if this difference is due to modeling individual and aggregate performance or a shift in strategy due to experimental conditions.

In contrast to the hypothesis testing model, the values reported for the associative model by Kachergis et al. (2012b) are largely consistent with our results. Their optimized values, fit to aggregate data, are $\alpha = 0.97$, $\chi = 0.05$, and $\lambda = 1.74$; values quite similar to the peaks of our distributions.

Finally, some model parameters strongly depend on the experimental condition. For example, the multimodal distribution of λ values (bottom right panel of Figure 4) suggests a mixture of different strategies across participants. We investigate this in Figure 5, which separates the best-fit λ values according to the ambiguity during training. It is evident that as learning conditions are more ambiguous, the λ value decreases. Since λ affects the weight assigned to novel words relative to familiar words, one interpretation of this is that the level of ambiguity during training has a strong impact on the extent to which novel items are emphasized during learning.

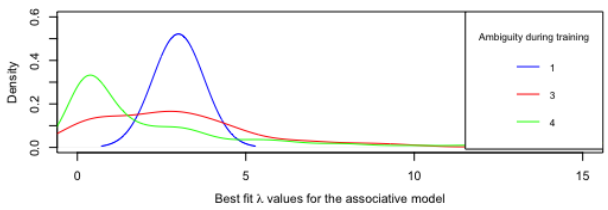

Figure 5: **Distribution of λ by ambiguity.** The best-fit λ values for the associative model (x axis) are plotted as a function of the level of ambiguity during training. The distribution of λ when there is no ambiguity (blue) has higher average values. As ambiguity increases (red, then green) the estimated λ values get smaller.

Discussion

In this work we investigated which of two computational models of cross-situational word learning offers a better account of word learning by individual participants across a wide range of conditions. For most people, the associative model (Kachergis et al., 2012a) outperforms the hypothesis testing model (Trueswell et al., 2013).

The advantage for the associative model is most pronounced in conditions in which the meaning of words is ambiguous during training, where it provides a better account for nearly all people. However, in conditions without ambiguity of word meaning, the hypothesis testing model outperformed the associative model for over 60% of participants. This advantage for the hypothesis testing model in unambiguous training conditions occurred for nearly every participant who experienced a uniform frequency distribution of words, but for participants in conditions with a Zipfian word frequency distribution the associative and hypothesis testing models provide the best account equally often.

The impact of other aspects of the learning environment on the relative performance of the two models was less striking. The total number of unique words present did not seem to influence which model was preferred, though there was some suggestion that the participants whose AIC was worst for the hypothesis testing model were in the conditions with the largest vocabulary size. Finally, manipulating if participants were required to guess during training had no effect on model preference as all relevant conditions were ambiguous during training and thus nearly all participants were best fit by the associative model.

Why the hypothesis testing model, despite multiple studies showing support for the model, performed consistently worse in the ambiguous learning contexts that require cross-situational learning is perhaps the biggest open question raised by these results. One possibility is that the hypothesis testing model, though designed to account for individual learning behavior, is not sufficiently flexible to account for the variation across participants. Restricting the memory strength to two possible values might provide a good account of aggregate data but be too rigid for matching individual behavior.

Another possible explanation for the worse performance of the hypothesis testing model is that even if people do form hypotheses about word-object pairs, they are also incorporating some co-occurrence information to shape their representations. This class of hybrid learning mechanisms, which incorporate both hypothesis testing and associative learning mechanisms (Yurovsky & Frank, 2015), provide a suggestion of additional types of models that might better capture the range of learning behavior in the ambiguous conditions. Similarly, Pursuit (Stevens et al., 2017), a recent variant of the Propose-but-Verify model that retains disconfirmed meanings and counts of referential success, might also improve on the performance of the earlier hypothesis testing model by finding a balance between testing hypotheses and gathering some co-occurrence information.

A final explanation of this effect may be due to specific aspects of the model fitting in this study. These choices include how the hypothesis testing model was extended to produce probability distributions across responses, the 10-fold simulation of parameter values to compute the median log likelihood, or the choice of AIC for model comparison instead of measures that have higher penalties for model complexity (e.g. BIC) or flexibility (Navarro, Pitt, & Myung, 2004).

Despite the clear advantage across many conditions for one model in this comparison, further work is clearly needed to fully understand the learning mechanisms and representations that underlie word learning. These include evaluating alternative models (e.g. Yu & Smith, 2012; Yurovsky & Frank, 2015; Stevens et al., 2017), expanding the range of evaluation techniques, and constraining models with additional data (e.g. Kachergis & Yu, 2018) or conditions (e.g. Hendrickson & Perfors, 2018).

References

Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, *19*(6), 716–723.

Aravind, A., de Villiers, J., Pace, A., Valentine, H., Golinkoff, R., Hirsch-Pasek, K., ... Wilson, M. S. (2018). Fast mapping word meanings across trials: Young children forget all but their first guess. *Cognition*, *177*, 177–188.

Aussems, S., & Vogt, P. (2018). Adults use distributional statistics for word learning in a conservative way. *IEEE Transactions on Cognitive and Developmental Systems*.

Hendrickson, A., & Perfors, A. (2018, Nov). *Cross-situational learning in a zipfian environment.* PsyArXiv. Retrieved from `psyarxiv.com/6jumv`

Horst, J., & Hout, M. (2015). The novel object and unusual name (NOUN) database: A collection of novel images for use in experimental research. *Behavior Research Methods*, *48*(4), 1393-1409.

Kachergis, G., & Yu, C. (2018). Observing and modeling developing knowledge and uncertainty during cross-situational word learning. *IEEE Transactions on Cognitive and Developmental Systems*, *10*(2), 227–236.

Kachergis, G., Yu, C., & Shiffrin, R. M. (2012a). An associative model of adaptive inference for learning word–

referent mappings. *Psychonomic Bulletin & Review*, *19*(2), 317–324.

Kachergis, G., Yu, C., & Shiffrin, R. M. (2012b). Cross-situational word learning is better modeled by associations than hypotheses. *IEEE Conference on Development and Learning*, 1–6.

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

Navarro, D. J., Pitt, M. A., & Myung, I. J. (2004). Assessing the distinguishability of models and the informativeness of data. *Cognitive Psych.*, *49*(1), 47–84.

Pinker, S. (1984). *Language learnability and language development*. Cambridge, MA: Harvard University Press.

Rasilo, H., & Räsänen, O. J. (2015). Computational evidence for effects of memory decay, familiarity preference and mutual exclusivity in cross-situational learning. In *CogSci.*

Reisenauer, R., Smith, K., & Blythe, R. (2013). Stochastic dynamics of lexicon learning in an uncertain and nonuniform world. *Physics Review Letters*, *110*(258701).

Smith, K., Smith, A. D., & Blythe, R. A. (2009). Reconsidering human cross-situational learning capacities: A revision to yu & smiths (2007) experimental paradigm. In *CogSci.*

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

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

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

Vogt, P. (2012). Exploring the robustness of cross-situational learning under Zipfian distributions. *Cognitive Science*, *36*, 726–739.

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

Woodard, K., Gleitman, L. R., & Trueswell, J. C. (2016). Two-and three-year-olds track a single meaning during word learning: Evidence for propose-but-verify. *Language Learning and Development*, *12*(3), 252–261.

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

Yu, C., & Smith, L. (2012). Modeling cross-situational word-referent learning: Prior questions. *Psych. Review*, *119*(1), 21–39.

Yurovsky, D., & Frank, M. C. (2015). An integrative account of constraints on cross-situational learning. *Cognition*, *145*, 53–62.