Memory Constraints on Cross Situational Word Learning

Christine Soh (sohc@sas.upenn.edu)
Department of Linguistics, 3401 Walnut Street
Philadelphia, PA 19104 USA

Charles Yang (charles.yang@ling.upenn.edu)
Department of Linguistics and Computer Science, 3401 Walnut Street
Philadelphia, PA 19104 USA

Abstract

A simple memory component is amended to local ("Pursuit"; Stevens, Gleitman, Trueswell, and Yang (2017)) and global (e.g., Yu and Smith (2007); Fazly, Alishahi, and Stevenson (2010)) models of cross-situational word learning. Only a finite (and small) number of words can be concurrently learned; successfully learned words are removed from the memory buffer and stored in the lexicon. The memory buffer improves the empirical coverage for both local and global learning models. However, the complex task of homophone learning (Yurovsky & Yu, 2008) proves a more decisive advantage for the local model (dubbed Memory Bound Pursuit; MBP). Implications and limitations of these results are discussed.

Keywords: memory; word learning; language acquisition; statistical learning; cross situational word learning; mutual exclusivity

Introduction

Understanding early word learning presents a great challenge, as many factors, both internal and external, affect the child's learning process. Multiple cognitive domains including attention, memory, and language abilities impact both word learning and experimental probing tasks. In the cross-situational word learning paradigm of experiments, the learner must be able to remember images and words, which are often novel; in the real world, the child must build and access their memory to use the words that they have learned (Vlach & DeBrock, 2017).

Nevertheless, the role of memory has not been systematically evaluated in computational models of word learning, which have primarily focused on the mechanisms of tracking and updating the word-meaning associations. Global learning models, to use a term from (Stevens et al., 2017), tabulate co-occurrence statistics of word-meaning pairs across learning instances (Yu & Smith, 2007; Fazly et al., 2010), which allows learners to use the entirety of their past experience in revising and developing their emerging lexicon. By contrast, hypothesis testing approaches allow learners to hold onto only a single hypothesized meaning locally for each word they encounter (Medina, Snedeker, Trueswell, & Gleitman, 2011). In its extreme form, a word is paired with only one hypothesized meaning at any time (Trueswell, Medina, Hafri, & Gleitman, 2013). The Pursuit model (Stevens et al., 2017) combines features of the local and global approaches. Like the global model, Pursuit allows for a word to be associated with multiple hypotheses, which are tracked across learning instances. Like the local model, Pursuit only evaluates a single hypothesis at any given time. In particular, only the best ranked hypothesis is tested ("pursued") and updated; lower ranked hypotheses, if they exist, are not considered at all, keeping the computational cost to a minimum.

Previous research (Stevens et al., 2017) shows that Pursuit offers broader coverage of experimental findings than both local and global models, including a paradigm study of (global) cross-situational word learning (Yu & Smith, 2007). However, several difficulties remain. First, despite the ability to maintain multiple hypotheses, it is not clear how Pursuit can capture the findings of homophone learning (Yurovsky & Yu, 2008). If the advantage of the best hypothesis over the alternatives is too large, then only the best will be pursued and learned. If the advantage is not decisive, then no hypothesis will emerge as the winner. Second, previous studies (Smith, Smith, & Blythe, 2011; Kachergis, Yu, & Shiffrin, 2012) have shown that massed presentation provides more favorable condition for interleaved presentation. Pursuit, however, has no way of distinguishing these learning conditions. Finally, Pursuit predicts that the best ranked hypothesis should always be the winning outcome, but that would result in learning accuracy exceeding the level exhibited by experimental subjects. To remedy this issue, Stevens et al. (2017) incorporated a recall parameter that controls the probability with which the best hypothesis is retrieved. This allows the model to fit most of the reported experimental results but at the expense of tuning a post-hoc free parameter.

There have been models in the global learning tradition that explore the effect of memory on cross-situational learning (Kachergis et al., 2012; Ibbotson, López, & McKane, 2018; Holehouse & Blythe, 2018). However, these models largely remain at the level of abstract simulation and have not been systematically tested against a wide range of behavioral results. In this paper, we introduce a simple memory component to complement word learning models. The key idea is that the tabulation of evidence for hypothesized meanings takes place in a working memory buffer, which can only hold a finite (and small) number of words (i.e., labels). The buffer functions as a queue with the least recently used word removed once it reaches capacity. However, if a meaning hypothesis for a word is learned by reaching some threshold (in a manner to be made clear) prior to its removal, it is shifted to the lexicon thereby emptying a slot in the buffer for additional words to be learned, including an additional meaning for the same word (i.e., homophone).

We first show that the memory buffer improves the empirical coverage of Pursuit as well as global learning models by presenting results from several behavioral experiments considered in (Stevens et al., 2017). However, the complex task of homophone learning (Yurovsky & Yu, 2008) proves a more decisive advantage for the Memory Bound Pursuit (MBP) model.[1] Implications and limitations of these results are then discussed.

Incorporating a Memory Constraint

We propose that word learning be modeled with two components: a *memory buffer* where hypothesized word meanings are held and updated, and a *lexicon* to which established word meanings are permanently stored. Crucially, the memory buffer is finite and quite small (Miller, 1956): only a limited number of words (i.e., labels) can be learned concurrently.

	Encountered word	Queue (size 4)			
1.	A				A
2.	C			A	C
3.	D		A	C	D
4.	A		C	D	A
5.	B	C	D	A	B
6.	D	C	A	B	D
7.	E	A	B	D	E

Figure 1: An example of the memory buffer in use. As words are encountered, they move to the top (on the right side), and when the queue is full, the least recently encountered word (on the left side) is forgotten, including its set of referent hypotheses and associations.

The Memory Buffer We implement a least recently used algorithm and update the buffer such that when a word is encountered, it moves to the front, as seen in Figure 1, where the front of the queue is the right side. The availability of human memory for particular items have been shown to correspond with the recency and frequency of prior exposures to the item. When the memory buffer is full, the least recently encountered word, taken from the back, is forgotten along with its associated referents. We believe that the memory buffer is the simplest computational mechanism (Belady, 1966) that embodies the size limit and other core properties of working memory (Ebbinghaus, 1913) and is strongly similar to several influential models (Waugh & Norman, 1965; Atkinson & Shiffrin, 1968; Anderson & Schooler, 1991).

MBP

Description of the Model MBP, based on the Pursuit model presented in Stevens et al. (2017), incorporates the memory buffer described above. When the model encounters a new word, it selects a referential hypothesis from the available objects. Following a (stochastic) variant of mutual exclusivity, objects already associated with other words/labels are less likely to be chosen. When a hypothesis is confirmed (i.e. the referential hypothesis is present in the set of available objects), the association value is increased, and when it is not confirmed, the model picks a new hypothesis at random from the options. This update strategy is shown in Figure 2. Crucially, only the best hypothesis is tested ("pursued") and the model does not consider lower ranked hypotheses even if the best fails.

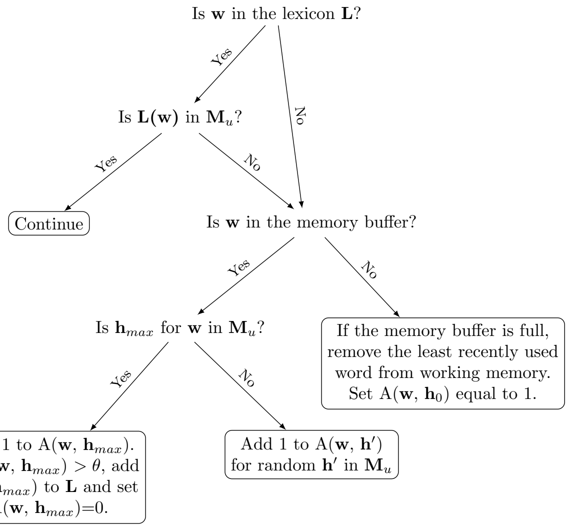
Is $\mathbf{w}$ in the lexicon $\mathbf{L}$?
Yes
No
Is $\mathbf{L}(\mathbf{w})$ in $\mathbf{M}_u$?
Yes
No
Continue
Is $\mathbf{w}$ in the memory buffer?
Yes
No
Is $\mathbf{h}_{max}$ for $\mathbf{w}$ in $\mathbf{M}_u$?
If the memory buffer is full, remove the least recently used word from working memory. Set $A(\mathbf{w}, \mathbf{h}_0)$ equal to 1.
Yes
No
Add 1 to $A(\mathbf{w}, \mathbf{h}_{max})$. If $A(\mathbf{w}, \mathbf{h}_{max}) > \theta$, add $(\mathbf{w}, \mathbf{h}_{max})$ to $\mathbf{L}$ and set $A(\mathbf{w}, \mathbf{h}_{max})$=0.
Add 1 to $A(\mathbf{w}, \mathbf{h}')$ for random $\mathbf{h}'$ in $\mathbf{M}_u$

Figure 2: A decision tree for MBP in the process of updating the association values in working memory. **w** stands for the word in question, **L** is the learned lexicon, $\mathbf{M}_u$ is the set of referents present in the utterance, $\mathbf{h}_{max}$ is the best hypothesis for a word, **A** is the set of associations between words and their hypothesized meanings, $\mathbf{h}_0$ is the initial hypothesis, which is selected by a principle of mutual exclusion, and θ is the threshold parameter.

MBP has a simplified mechanism for updating the association values for the hypotheses. We reward linearly when a hypothesis is confirmed, keeping a count of co-occurrences. For small numbers of instances, as we have with the experimental simulations, this simplification is approximately the same as the reinforcement learning style probabilistic adjustment used in the original Pursuit model. If the memory buffer is full when a new word is encountered, then the least recently encountered word is forgotten, along with its set of associations. The meaning for a word is learned if its score exceeds a threshold. Importantly, the threshold value is not absolute but reflects the competitive nature of learning: a meaning wins only if its score is sufficiently higher than its competitors. This is implemented in most computational models by score renormalization and smoothing. For simplicity, we assume

[1] https://github.com/christinesoh/memory-bound-pursuit

that if the score of a meaning is at least twice that of its closest competitor, the meaning is learned and shifted to the lexicon. Once a meaning is learned, the word may still remain in the buffer (until it is pushed out by new words), allowing for the learning of a second meaning in the case of homophony.

MGX

Description of the Model A similar modification was made to an implementation of the Fazly et al. (2010) global cross-situational (GX) learning model, dubbed the Memory Buffer Global Cross-situational (MGX) model. In particular, we use a modification of the GX model; see Stevens et al. (2017) for details. The modification was necessary, for otherwise the target hypothesis would be learned 100% of the time which is clearly at odds with human behavior in word learning experiments. In GX, all word-meaning associations available in each learning instance with no restriction on memory. MGX adds a memory component to GX similar to the MBP amendment of Pursuit. In MGX, if the memory buffer is full when a new word is encountered, then the least recently encountered word is removed from the buffer. With each learning instance, the emerging lexicon is updated by adding word-meaning pairs that have an association value greater than the threshold value. The learned word remains in the working memory until it is forgotten, maintaining the advantage and key characteristic of cross-situational learning that all remembered previous encounters are used when generating the lexicon – unless the buffer size limit is exceeded.

Testing Models on Experimental Conditions

At the end of learning, both MBP and MGX will have a lexicon of learned words as well as word-meaning associations in the memory buffer that have not yet reached the learning threshold. Most word learning studies provide a test immediately after training by asking the subject to select the referent of a word among a set of non-target alternatives: it is thus reasonable to assume that the content of the memory buffer is still accessible. In testing, the model first checks its lexicon: if the word is in the lexicon and the learned referent is a possible option, it selects the learned referent. Next, the model checks its memory buffer: if the word is in the memory buffer, then it samples from the options weighted by the association value, following Luce's choice axiom. Finally, if the word in question is neither in the lexicon nor in the set of associations, the model selects randomly from the options.

For each of the experiments, we set the mean size of the experimental subject's memory buffers to 10; we return to the choice of the memory buffer size in the Conclusion. To account for individual variation in memory capacity across subjects, we sample from a normal distribution with the mean centered at an value with a standard deviation of 1. Because MBP randomly selects its hypotheses, and because the multiple choice selection has stochastic behavior, the model is run 300 times with the accuracy averaged across the runs. Unlike its GX predecessor, which has infinite memory and thus produces deterministic output for any learning sequence, the variation in memory buffer size under MGX produces a non-deterministic output. Thus MGX is also run 300 times.

The Reporting of Experimental Results For each of the experiments, the accuracies of each model in each given condition are presented, with the corresponding 95% confidence interval (CI) in parentheses. The values are bolded if the CI of the model's accuracy overlaps with the observed CI, indicating that the model is behaving like the human participants, whose performance is noted as "Reported."

Yu and Smith (2007)

Experimental Setup This experiment provided key evidence for cross-situational word learning across a series of referentially ambiguous learning instances. The adult participants in Yu and Smith (2007) were exposed to learning trials with 2, 3, or 4 novel words and the matching number of novel referents, where with increased ambiguity, the participants' accuracy decreased.

Experimental Results As with the original Pursuit model, MBP has overlapping CI with the reported results, as shown in Table 1. Additionally, MGX marks a considerable improvement over the original GX model, highlighting the benefits of incorporating a memory component.

	4x4	3x3	2x2
Reported	0.53 (0.37-0.69)	0.76 (0.62–0.90)	0.89 (0.79–0.99)
Pursuit	**0.71** **(0.62-0.80)**	**0.84** **(0.76–0.91)**	**0.96** **(0.92–0.99)**
GX	0.96 (0.95-0.97)	0.97 (0.96-0.98)	**0.99** **(0.99-1.00)**
MBP	**0.43** **(0.41-0.44)**	**0.57** **(0.55-0.59)**	**0.77** **(0.75-0.79)**
MGX	**0.58** **(0.57-0.60)**	**0.69** **(0.67-0.70)**	**0.81** **(0.79-0.83)**

Table 1: Accuracies at testing of each of the models in each of the conditions, presented in order of decreasing ambiguity. The bold indicates that the 95% confidence interval (CI) of the accuracy overlaps with the reported results' CI.

Trueswell et al (2013)

Experimental Setup Adult subjects learned on a sequence of trials in which they heard one nonsense word and five objects, guessing the object that the word was associated with at each trial. The authors tracked the subjects' responses over time and found that they were more likely to guess correctly when they had previously guessed the word correctly.

Experimental Results For both conditions, MBP and MGX produce confidence intervals overlapping for experimental results, a considerable improvement over their respective predecessors.

	Previous Correct	Previous Incorrect	Significant Difference
Reported	0.47	0.21	Yes
	(0.33–0.70)	(0.17–0.25)	
Pursuit	0.80	**0.21**	Yes
	(0.78–0.83)	**(0.20–0.22)**	
GX	0.87	0.65	Yes
	(0.86-0.88)	(0.64-0.67)	
MBP	**0.63**	**0.21**	Yes
	(0.61-0.65)	**(0.20-0.22)**	
MGX	**0.57**	**0.26**	Yes
	(0.55-0.59)	**(0.25-0.27)**	

Table 2: Accuracies of each of the models given that the previous guess of the word was correct or not. The bold indicates that the CI of the accuracies overlaps with those reported. The column 'Significant Difference' shows whether there is a significant difference between the two conditions (Previous Correct and Previous Incorrect).

	AAAPPP	APAPAP	PAPAPA	PPPAAA
Reported	0.16	0.23	0.27	0.34
	(0.09-0.25)	(0.16-0.32)	(0.20-0.34)	(0.25-0.43)
Pursuit	**0.17**	**0.18**	**0.31**	**0.34**
	(0.13-0.19)	**(0.14-0.20)**	**(0.27-0.35)**	**(0.31-0.37)**
GX	0.90	0.90	0.91	0.91
	0.88-0.92	0.88-0.92	0.89-0.93	0.88-0.92
MBP	**0.26**	**0.28**	0.37	**0.45**
	(0.24-0.27)	(0.26-0.30)	(0.35-0.39)	(0.43-0.47)
MGX	**0.20**	0.38	0.65	0.83
	(0.18-0.22)	(0.36-0.39)	(0.63-0.67)	(0.81-0.85)

Table 3: Accuracies at testing of each of the models in each of the conditions, presented in order of increasing accuracy for human participants. The value is bolded for the model's performance if the 95% CI overlapped with the reported results' 95% CI for accuracy. Note that MBP generally captures the trend of higher performance in the PPPAAA condition, and the MGX does as well (while GX fails to).

Koehne et al (2013)

This experiment suggests that word learners can maintain multiple possible meanings for a word, with differences in presentation order affecting the learning of the meaning.

Experimental Setup In a similar experimental setup to the Trueswell et al. (2013) experiments, adult participants hear a novel word with four objects. The crucial detail is that each word is associated with two referents, one which appears every time (the hundred percent referent, or HPR), and one which appears 3 of the 6 times the word is heard (the fifty percent referent, or FPR). There are four conditions that the participants are split into, corresponding to the order of which the FPR is present (P) and absent (A) when the word is seen: PPPAAA, PAPAPA, APAPAP, and AAAPPP. In testing, only the FPR is present as an option.

Experimental Results The Pursuit model replicated the reported trends; the consistent initial evidence from the first three trials in the PPPAAA condition resulted in a higher likelihood of giving the FPR a high probability score. This study provided evidence that human learners likely were able to store more than just a single hypothesis, as they performed above chance in the PAPAPA and PPPAAA conditions, where the FPR was not present in the final trial. Similarly, MBP gives an advantage to the FPR in the PPPAAA condition. The memory constraint allows MGX to capture the trend that its original model could not: having the FPR present in the first trial creates an advantage for learning, and the PPPAAA condition results in maximal performance. That is, GX performed similarly across the four conditions, while MGX captures the differences across the conditions.

Yu and Yurovsky (2008)

Experimental Setup (Yurovsky & Yu, 2008) tested the learning of homophones and explored the how the mode of presentation – massed or interleaved – affected learning performance. In this study, the subjects were presented with a series of 27 trials. Each trial consisted of 4 words and 4 objects without any information as to the word-meaning pairings. There were 6 double-meaning words, corresponding to homophones, and 6 single-meaning words, and each word-meaning pair was presented 6 times. Individual trials were ambiguous in this way, but words always co-occurred with their correct referents.

The key manipulation in this study was the order in which word-referent pairs were presented. In Experiment 1, the training was split into two halves: the first half had one set of word-referent pairings, and in the second half, the same words co-occurred with new referents, while the old referents were absent. That is, in the first six occurrences of word 'A', the referent 'a1' appeared, while the last six occurrences of the word 'A' co-referred with meaning 'a2.' This condition is called the "Mass" condition, since the meanings appeared in two masses. Experiment 2, the "Interleaved" condition, presented the double words' meanings in an alternating order; that is, after the 'A'-'a1' pair is seen, the next occurrence of the word 'A' co-occurs with the referent 'a2.' The training data was controlled for in that pairs of words were presented in trials an equal number of times.

Additionally, there was neither an indication that some words map to two referents instead of one nor information as to which are homophones vs. single meaning words. For evaluation, each of the trained words appeared with 4 referents in a random order. There were three conditions for testing that corresponded to the possible options for the double words: in the *Primacy*, or First Meaning condition, the first referent for the double word appeared with 3 other referents; in the *Recency*, or Second Meaning condition, the second was present, and in the *Both* condition, both referents were present.

Experimental Results Yurovsky and Yu present experimental results run on 48 adult participants, shown in Figure 3. There are two key results: (1) there is a bias for the first meaning in the mass condition; (2) the performance of the single words dropped between the mass and the interleaved experiments. While single words are learned better than double words in the mass condition, there is not a statistically significant difference between accuracies on single words and double words in the interleaved condition. Additionally, the participants performed above chance for all the conditions, indicating that word learning was occurring. The increased number of total word-meaning pairs that need to be learned concurrently (18 in the interleaved condition, and 12 in the massed condition) increase the processing and memory load, causing a decrease in performance. It is clear that neither Pursuit nor GX could account for these findings.

MBP and MGX Results The experimental results for MBP and MGX are presented in Figure 3. As with the experimental results, both models perform above chance for each of the test conditions in both experiments.

MBP captures both key results, but MGX shows less sensitivity to the temporal structure of learning. First, we see a bias for the First Meaning referent with MBP in Exp 1 (mass) that we do not see in Exp 2 (interleaved). This is seen in the Primacy Test and Recency Test conditions, as well as in the distribution of responses for the Both Test condition. MGX shows the bias in the distribution of responses of the Both Test, but not in the other test conditions. Creating the lexicon incrementally allows for a stronger primacy effect, as the first meaning can be learned and committed to memory if presented in a massed order. Second, the performance of the single words decreases between the massed and interleaved for MBP but not for MGX. For MBP, the single words are learned at a slightly higher rate than the reported, but the combined responses in the Both test condition exceed the performance of the single words.

The introduction of a memory buffer thus provides a mechanism to learn homophones. In general, one meaning is learned first and removed from the buffer to the lexicon before a second meaning can emerge. Both MBP and MGX learn single word-meaning pairs better than word-meaning pairs for double words. Yurovsky and Yu (2008) reported that in Exp 2, there was no significant difference between single and double words; however, this is not completely captured in the models, as the interleaved condition presents much less opportunity for the model to confirm a meaning for a double word than for a single word.

To summarize, we have presented evidence that a simple memory buffer considerably enhances the empirical coverage of both local and global word learning models. However, the complex homophone learning task (Yurovsky & Yu, 2008) suggests that all things being equal, a local hypothesis testing model such as Pursuit (MBP) provides the best overall fit with the empirical data.

Conclusion

We have shown that a constraint that limits the size of the working memory and keeps only the most recently encountered words may be a strong starting point for the integration of memory into cross-situational word learning models. In a nutshell, learning can only succeed if the word stays in the working memory long enough to reach the critical threshold of clearance. Equivalently, learning also succeeds if accumulation of confirming evidence takes place fast enough before the word is pushed out of the working memory, which accounts for the advantage of massed presentation of learning instances.

The MBP model is extremely simple and contains only two parameters: memory buffer size and a threshold for deciding whether a meaning is sufficiently supported against its competitors (and thus "learned"). Both parameters correspond to the psychological reality of human subjects (e.g., individual's working memory size does vary), and there are no additional free parameters to tune. In our studies, we have set the mean memory buffer size to 10, a value similar to previous estimates of human information processing capacity (Miller, 1956). Variation around this mean is able to account for a wide range of findings under different experimental conditions (e.g., real vs. novel objects, the number of words to be learned, the degree of ambiguity in each learning instance, the interval between successive presentation of words). While additional apparatus will no doubt improve empirical coverage, we propose that MBP be a baseline model for word learning due to its simplicity.

Several lines of future research suggest themselves. First, MBP can be applied to more realistic word learning studies (Gillette, Gleitman, Gleitman, & Lederer, 1999) where the ambiguity of each learning instance can be individually manipulated and assessed. Preliminary results using the stimuli from (Medina et al., 2011) shows that MBP captures the trajectory of the subject's hypotheses as learning instances are incrementally presented, revealing a previously unnoticed effect of individual's memory capacity on the learning outcome. Second, MBP naturally leads to within-subject studies in which the participant's memory buffer size can be independently assessed so as to better understand individual variability in word learning. Finally, the memory buffer can offer a model of development: the memory buffer size increases from childhood to adulthood. It is uncontested that adults perform much better than children overall but children behave similarly to adults with regards to the effect of increased ambiguity in context to performance (Suanda, Mugwanya, & Namy, 2014). Running MBP with a decreased memory buffer size of 4, we find that the performance does indeed decrease while the trend of decreased performance with increased accuracy remains.

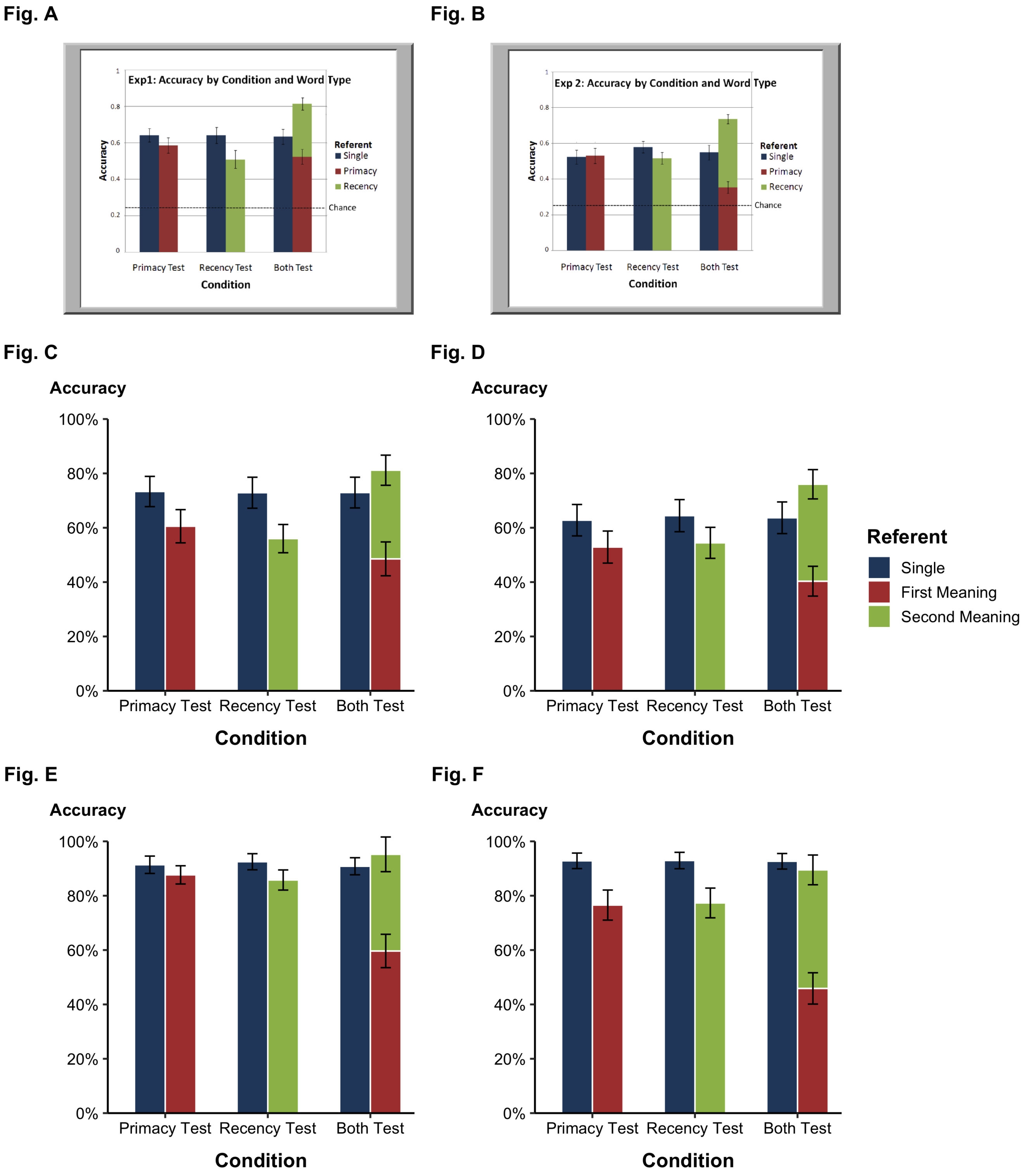

Figure 3: Simulations of Yu and Yurovsky (2008) Experiments 1 and 2. Figures A and B show the accuracy of human subjects, as reported in Yu and Yurovsky (2008); C and D show the accuracy of MBP; E and F show the accuracy of MBX.

Acknowledgments

We thank Ryan Budnick, June Cho, Lila Gleitman, Alice Healy, Alexander LaTourette, Kathryn Schuler, Jon Stevens, and John Trueswell for helpful discussions about this work.

Addendum

The model has been further developed to give a stronger developmental account. Instead of deleting the least recently encountered word, the model has been modified to delete a random word. This aligns more with the modal model of memory and does not impact the results shown here, as the memory buffer is large enough for the adult experimental participants that there is no significant effect.

References

Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological science*, *2*(6), 396–408.

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In *Psychology of learning and motivation* (Vol. 2, pp. 89–195). Elsevier.

Belady, L. A. (1966). A study of replacement algorithms for a virtual-storage computer. *IBM Systems journal*, *5*(2), 78–101.

Ebbinghaus, H. (1913). Memory (ha ruger & ce bussenius, trans.). *New York: Teachers College.(Original work published 1885)*, *39*.

Fazly, A., Alishahi, A., & Stevenson, S. (2010). A probabilistic computational model of cross-situational word learning. *Cognitive Science*, *34*(6), 1017–1063.

Gillette, J., Gleitman, H., Gleitman, L., & Lederer, A. (1999). Human simulations of vocabulary learning. *Cognition*, *73*(2), 135–176.

Holehouse, J., & Blythe, R. A. (2018). Cross-situational learning of large lexicons with finite memory. *arXiv preprint arXiv:1809.11047*.

Ibbotson, P., López, D. G., & McKane, A. J. (2018). Goldilocks forgetting in cross-situational learning. *Frontiers in psychology*, *9*, 1301.

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.

Medina, T. N., Snedeker, J., Trueswell, J. C., & Gleitman, L. R. (2011). How words can and cannot be learned by observation. *Proceedings of the National Academy of Sciences*, *108*(22), 9014–9019.

Miller, G. A. (1956). The magic number seven plus or minus two: Some limits on our capacity for processing information. *Psychological review*, *63*, 91–97.

Smith, K., Smith, A. D., & Blythe, R. A. (2011). Cross-situational learning: An experimental study of word-learning mechanisms. *Cognitive Science*, *35*(3), 480–498.

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

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.

Trueswell, J. C., Medina, T. N., Hafri, A., & Gleitman, L. R. (2013). Propose but verify: Fast mapping meets cross-situational word learning. *Cognitive psychology*, *66*(1), 126–156.

Vlach, H. A., & DeBrock, C. A. (2017). Remember dax? relations between children's cross-situational word learning, memory, and language abilities. *Journal of memory and language*, *93*, 217–230.

Waugh, N. C., & Norman, D. A. (1965). Primary memory. *Psychological review*, *72*(2), 89.

Yu, C., & Smith, L. B. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological science*, *18*(5), 414–420.

Yurovsky, D., & Yu, C. (2008). Mutual exclusivity in cross-situational statistical learning. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 30).