A large-scale comparison of cross-situational word learning models

Anonymous CogSci submission

Abstract

One problem language learners face is extracting word meanings from scenes with many possible referents. Despite the ambiguity of individual situations, a large body of empirical work shows that people are able to learn cross-situationally when a word occurs in different situations. Many computational models of cross-situational word learning have been proposed, yet there is little consensus on the main mechanisms supporting learning, in part due to the profusion of disparate studies and models, and lack of systematic model comparisons across a wide range of studies. This study compares the performance of several extant models on a dataset of 44 experimental conditions and a total of 1,696 participants. Using cross-validation, we fit multiple models representing theories of both associative learning and hypothesis-testing theories of word learning, find two best-fitting models, and discuss issues of model and mechanism identifiability. Finally, we test the models’ ability to generalize to additional experiments, including developmental data.

Keywords: cross-situational word learning; model comparison; language learning

Introduction

Learning word meanings is a crucial aspect of the process of language acquisition. If a learner’s task is to connect the words they hear to the potential referents and eventually to generalizable meanings, the challenge appears formidable (Quine, 1960): words are heard in rich and cluttered contexts. Yet children learn words very quickly. How do they do this?

An array of social (Yurovsky & Frank, 2017), pragmatic (Clark, 1987: Markman, 1992), and linguistic (Gleitman, 1990) cues are often available to reduce a learner’s uncertainty. Yet at its heart, these information sources can be seen as reducing the uncertainty inherent in the task of cross-situational mapping (Frank, Goodman, & Tenenbaum, 2009): that is, cross-referencing which refersent a word has consistently appeared with across multiple uses (Carey & Bartlett, 1978; Gleitman, 1990). If learners can learn cross-situationally, they might be able to pool information about reference (and hence word meaning) over time and learn even absent perfectly disambiguated input in any particular situation. Can children (or even adults) learn in this way?

The general idea of cross-situational word learning has been adopted into a variety of experimental paradigms for testing the word learning ability of infants (e.g., Smith & Yu, 2008; Vlach & Johnson, 2013), children (e.g., Akhtar & Montague, 1999; Suanda, Mugwanya, & Namy, 2014; Vlach & Sandhofer, 2012), and adults (e.g., Yu & Smith, 2007; Smith, Smith, & Blythe, 2011). In such experiments, participants are typically presented with a series of training trials comprised of multiple possible referents (e.g., an array of 2-4 novel objects) and one or more spoken nonce words, and (when possible) instructed to learn the meaning(s) of the presented words. These training trials typically present 10–20 distinct word-referent pairs to be learned across 10–50 trials, presenting each word/referent 3–10 times across a total span of < 10 minutes. Learning is then typically assessed by presenting each word and asking learners to make a forced choice of the best referent, either from the entire set or from a small subset of referents.

People of all ages are able to learn word-object mappings in such paradigms at a level significantly above chance, but their performance is typically far from perfect in all but the most trivial versions of the task (Yurovsky & Frank, 2017). Numerous training manipulations have been found to influence performance. How often mappings are presented (Kachergis, Yu, & Shiffrin, 2016; Suanda & Namy, 2012), the interval between appearances of a mapping (Vlach & Johnson, 2013; Vlach & Sandhofer, 2012), and the number of mappings presented per trial (Yu, 2008) all have been shown to affect learning performance. In sum, the evidence is strong that people can learn word cross-situationally: we are somehow able to integrate evidence (co-occurrence of words and objects) across trials and use that evidence to learn and remember new word meanings in cross-situational learning experiments. But what are the psychological mechanisms supporting this process? And critically, do these mechanisms support naturalistic word learning?

Computational models provide one strategy for instantiating proposals about mechanism. Such models can then be applied to naturalistic data to make a first assessment of whether cross-situational learning could be a viable strategy for word learning “in the wild”. Accordingly, there has been great interest in creating and testing computational models of this task. These models often represent distinct intuitions of how learning operates, variously representing the problem as logical inferencing (e.g., Siskind, 1996), proposing and testing hypotheses (e.g., Trueswell, Medina, Hafri, & Gleitman, 2013), or accumulating graded associations (e.g., Fazly, Alishahi, & Stevenson, 2010; Kachergis, Yu, & Shiffrin, 2012).

It is unknown which of these models – or even which broad class of models – best fits the breadth of experimental
data. While most models of cross-situational learning have typically been tested against at least a few datasets, to our knowledge there has been no systematic comparison of models across a wide variety of experimental conditions. Identifying the model(s) that best predict human performance is thus an important goal for consolidating this broad literature. More broadly, identifying such models could be an important step in assessing the contribution of cross-situational learning to word learning more generally.

Thus, the goal of our current study is to test the ability of a range of cross-situational word learning models in accounting for a wide range of experimental data. Our goal is to scale up both the number of models evaluated and the number of experiment designs and participants being used for evaluation. Different models often contain a range of parameters that must be fit to data, an issue that has hindered previous comparison work, either because the parameters have been allowed to vary freely across multiple conditions, or because models’ performance is compared on different datasets. Thus, we focus on out-of-sample prediction of human data using cross-validated parameter fits. We end by examining results from generalization of models to other experiments not included in the initial evaluation and parameter fitting.

**Method**

**Procedure**

Our goal is to fit a set of models to a set of datasets; members of each of these sets are listed below. Each model typically is defined as a function that takes input data and a group of parameter values (often between 2–4) and then can be evaluated on test trials (typically multi-alternative forced choice test). Each model then returns predictions on each test trial, which can be scored against the “correct” (intended by the design) answer and averaged to produce an analogue to human performance. Our primary outcome measure is the correlation between each model’s performance for each item in a specific experimental condition and item-level human performance on that same experimental condition.

Although input data and test trials are features of specific experimental tests, each model has parameters that must be specified. Typically the settings of these parameters are critical to the predictive performance of the model. Hence, to ensure a fair comparison between models, these parameters must be fit to data. In all cases, we optimize parameter values to minimize the sum of squared error (SSE) between model and human performance, weighting the contribution of each condition’s SSE by the number of participants in that condition. We optimize parameters using differential evolution via the DEoptim R package (Mullen, Ardia, Gil, Windover, & Cline, 2011). For the stochastic models, each parameter setting was evaluated using a simulated sample of 200 learners.

We conducted two evaluations of models. First, we optimize each model’s parameters on 80% of data and evaluate on the remaining 20% (5-fold cross-validation). Second, in the group fit regime we fit each model to all conditions and datasets and evaluate across all data. Though slightly overfit, this evaluation allows us to make the best test of generalization to other experiments from the literature.

**Models**

We fit a baseline co-occurrence model along with six associative models, two hypothesis-testing models, and two hybrid models that store multiple, graded hypotheses—only a subset of all possible associations. These models and their free parameters are briefly described below.

**Baseline Co-occurrence Model**

The baseline model simply tallies the co-occurrences of any words and objects appearing together on each trial, accumulating the counts in a word \( \times \) object matrix \( M \). At test, for word \( w_1 \) this model selects the correct referent \( o_1 \) according to Luce choice from the co-occurrences of \( w_1 \) with all referents: \( P(o_1|w_1) = M_{w_1,o_1}/\sum M_{w_1,o} \). The model’s probability of correctly choosing each word per condition is what is scored against people’s average probability of choosing the correct item.

**Probabilistic Associative Model (PA)**

The probabilistic associative model (Fazly et al., 2010) represents the meaning of each word \( w \) as a probability distribution \( p(w|o) \) over the objects appearing across the trials, and incrementally learns these distributions trial-by-trial. The association between a given \( w \) and \( o \) grows more quickly if \( p(w|o) \) is already high (a familiarity bias), unless some other word has a strong association with \( o \); thus, associations are competitive. A parameter \( \lambda \) is a small smoothing constant added to both the numerator and denominator of each \( p(w|o) \), further adjusted in the denominator by a \( \beta \) parameter representing the upper bound on the number of symbol types. At test, the correct referent for each word \( w \) is chosen according to Luce choice from the probability distribution over the objects that appeared with that word \( p(w|o) \).\(^\dagger\)

**Familiarity- and Uncertainty-biased Model (FU)**

The familiarity- and uncertainty-biased associative model (Kachergis et al., 2012) assumes that learners associate all words with all visible objects to some extent, but that they selectively attend more to some of the presented word-object pairings. Specifically, greater attention is directed to pairings that have previously co-occurred (a familiarity bias), but is also directed to stimuli that have no strong associates, tracked via the entropy \( (H) \) of their association strengths (e.g., novel stimuli, or stimuli that have diffuse associations with several stimuli). Thus, the model grows its association matrix as it experiences each trial, dynamically apportioning a fixed amount of associative weight (learning rate \( \chi \)) among the possible word-object associations, with the relative weight of the familiarity bias and the uncertainty bias determined by parameter \( \lambda \). Association strengths decay at a rate controlled by parameter \( \alpha \). The update rule for adjusting the association \( M_{w,o} \) between a given word \( w \) and object \( o \) on a given trial is:

\[ M_{w,o} = M_{w,o} + \lambda \underbrace{p(w|o) \cdot \left[ 1 - \left( 1 - \alpha \chi \right)^t \right]}_{\text{learning}} - \lambda \underbrace{p(w|o) \cdot \left( 1 - \left( 1 - \alpha \chi \right)^t \right)}_{\text{forgetting}} \]

\[ \text{where } t \text{ is the trial number.} \]

\( \dagger \) The original model implemented a 0-thresholded lexicon that discretized the probability matrix, but we found that including this mechanism resulted in a worse fit to the data.
\[ M_{w,o} = \alpha M_{w,o} + \frac{\chi \cdot e^{\lambda (H(w) + H(o))} \cdot M_{w,o}}{\sum_{w \in W} \sum_{o \in O} e^{\lambda (H(w) + H(o))} \cdot M_{w,o}} \]  

(1)

At test, given word \( w \) learners use Luce choice, choosing referent \( o \) from the \( m \) given alternatives in proportion to associative strength \( M_{w,o} \).

We also fit a stochastic sampling (FUs) version of this model (see Kachergis & Yu, 2017), which uses the same update equation, but samples only a single presented object for each word on a trial instead of updating all possible associations. This FU sampling model can still build multiple graded associations for each word (or object), but on any given run will accumulate only a sparse, randomized version of what the full associative model would build.

Strength-, Uncertainty-, & Novelty-biased (Str,Unc,Nov) Models We also test three variations of the FU model that use only a subset of the mechanisms of the full model in order to understand the contributions of each mechanism. The strength model lacks the uncertainty bias term, and thus only implements a bias for familiar (i.e., already-strong) associations, with learning rate (\( \chi \)) and decay \( \alpha \) parameters.

The uncertainty model lacks the strength bias term, and thus only implements a bias for stimuli with uncertain (high-entropy) associations. The novelty model also lacks the strength bias term, and substitutes novelty for the entropy terms (e.g., \( 1/(\text{frequency}(w)+1) \) instead of \( H(w) \)). These models have all three parameters of the original FU model, and operate in the same way as that model at test.

Bayesian Decay Model (BD) This previously-unpublished model updates the \( p(w|o) \) and the joint probability \( p(w,a) \) from trial to trial according to a likelihood function that reinforces the association between \( w \) and \( o \) when they co-occur (scaled by parameter \( \delta \)), and penalizes all associations that are not occurring on the trial (scaled by parameter \( \alpha \)). Thus, in contrast to other incremental associative learning models considered here (e.g., Fazly and the Kachergis class of models), this model globally updates the entire association matrix on each trial – both co-occurring pairs and non-co-occurring pairs, and re-normalizes the tracked probabilities at each time step. At test, this model uses a softmax choice rule with parameter \( \tau \) to select the best referent for each word from the presented alternatives.

Rescorla-Wagner Model (R-W) This associative model is inspired by the prediction-error-based learning model of Rescorla & Wagner (1972). Objects on each trial serve as cues to predict which words will be heard. When a given word is heard, the associations between that word and each object on the trial are scaled at a learning rate \( \beta \) in proportion to the magnitude of the difference between the prediction and the maximum association value (\( \lambda \)). Each trial, the association matrix is subject to decay (parameter \( \alpha \)). At test, this model uses Luce choice to select the best referent for each word from the presented alternatives.

Propose-but-Verify Model (PhV) In the propose-but-verify hypothesis testing model (Trueswell et al., 2013), a presented referent is selected at random for any presented word that has no remembered referent. The next time that word occurs, the previously-proposed referent is remembered with probability \( \alpha \), a free parameter. If the remembered referent is verified to be present, the future probability of recalling the word is increased by parameter \( \epsilon \). If the remembered referent is not present, the old hypothesis is assumed to be forgotten and a new proposal is selected from the available referents. This model implements trial-level mutual exclusivity by selecting new proposals only from among the referents that are not yet part of a hypothesis.

Guess-and-Test Model (GnT) The guess-and-test hypothesis testing model is based on the description given by Medina, Snedeker, Trueswell, & Gleitman (2011) of a one-shot (i.e. “fast mapping”) learning, which posits that “i) learners hypothesize a single meaning based on their first encounter with a word, ii) learners neither weight nor even store back-up alternative meanings, and iii) on later encounters, learners attempt to retrieve this hypothesis from memory and test it against a new context, updating it only if it is disconfirmed.” We give this model two free parameters: a probability of successful encoding (\( s \), hypothesis formation), and a probability \( f \) of forgetting a hypothesis at retrieval. At test, for each word the model chooses the currently hypothesized referent.

Pursuit Model (Pur) The pursuit model (Stevens, Gleitman, Trueswell, & Yang, 2017) is a hybrid model, potentially storing graded associations involving a given word and multiple referents (or vice-versa), while on any given trial greedily pursuing the strongest association for each presented word. If the strongest associated referent for a given word \( w \) is present on the trial, the association is strengthened at a rate determined by \( \gamma \). If the strongest referent for \( w \) is not present, the association is weakened (scaled down by \( 1 - \gamma \)) and the association with a random other available referent is strengthened. Novel words are given an initial association of strength \( \gamma \) with the available referent whose strongest association is the smallest, implementing an uncertainty bias in forming associations for novel words. A given word-object association only enters the lexicon if \( p(o|w) > \theta \), a threshold parameter. At test, this model chooses the best referent for each word by Luce choice from the lexicon, which is not necessarily 1-to-1.

Data

The modeled data are average accuracies from 726 word-object pairs in 44 experimental conditions, in which a total of 1696 subjects participated. A table containing details of each condition is available on OSF. The number of trials per condition ranged from 18 to 108, with 1-4 words and objects presented per trial, and a total of 12-24 pairs to be learned per condition. The bulk of these data have been previously published: data from 13 of the conditions were reported in Kachergis & Yu (2013), data from 12 conditions are from

\footnote{Information about the dataset: [https://osf.io/REDACTED]}
Table 1 shows the sum of squared error (SSE) and correlation (r) for each model's best-fitting parameters vs. average human performance on the 726 items, both for the CV fits and for the all-condition fits, along with the number of fitted parameters in each model (P). The results of the two fitting procedures mostly yielded similar SSEs and correlations, and a consistent rank-ordering of the models, with the Kachergis sampling model performing best, followed closely by the Fazly et al. model and then the associative Kachergis et al. model. The hypothesis testing models (Propose-but-Verify and Guess-and-Test), the Pursuit model, and R-W fit roughly as well as the 0-parameter baseline co-occurrence counting model, which had SSE = 32.0 and r = 0.53. Correlations between data and models are presented in Figure 1.

<table>
<thead>
<tr>
<th>Model</th>
<th>CV SSE</th>
<th>r</th>
<th>All SSE</th>
<th>r</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>FU</td>
<td>17.74</td>
<td>0.74</td>
<td>17.81</td>
<td>0.74</td>
<td>3</td>
</tr>
<tr>
<td>PA</td>
<td>18.72</td>
<td>0.73</td>
<td>18.77</td>
<td>0.73</td>
<td>2</td>
</tr>
<tr>
<td>FU</td>
<td>18.92</td>
<td>0.72</td>
<td>19.07</td>
<td>0.72</td>
<td>3</td>
</tr>
<tr>
<td>Nov</td>
<td>22.02</td>
<td>0.67</td>
<td>22.01</td>
<td>0.67</td>
<td>3</td>
</tr>
<tr>
<td>Unc</td>
<td>23.06</td>
<td>0.69</td>
<td>23.02</td>
<td>0.69</td>
<td>3</td>
</tr>
<tr>
<td>BD</td>
<td>23.95</td>
<td>0.62</td>
<td>23.71</td>
<td>0.63</td>
<td>3</td>
</tr>
<tr>
<td>Grn</td>
<td>31.49</td>
<td>0.50</td>
<td>30.98</td>
<td>0.51</td>
<td>2</td>
</tr>
<tr>
<td>R-W</td>
<td>31.80</td>
<td>0.62</td>
<td>31.79</td>
<td>0.62</td>
<td>3</td>
</tr>
<tr>
<td>Str</td>
<td>39.05</td>
<td>0.46</td>
<td>39.07</td>
<td>0.46</td>
<td>2</td>
</tr>
<tr>
<td>Pur</td>
<td>41.68</td>
<td>0.57</td>
<td>40.56</td>
<td>0.59</td>
<td>3</td>
</tr>
<tr>
<td>PbV</td>
<td>42.92</td>
<td>0.35</td>
<td>30.76</td>
<td>0.51</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 2: Correlations of models’ predictions.

<table>
<thead>
<tr>
<th>Model</th>
<th>Koehne</th>
<th>Medina</th>
<th>Suanda</th>
<th>Smith &amp; Yu</th>
<th>Adult RMSE</th>
<th>Dev RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>R-W</td>
<td>0.09</td>
<td>0.11</td>
<td>0.11</td>
<td>0.17</td>
<td>0.2</td>
<td>0.28</td>
</tr>
<tr>
<td>Nov</td>
<td>0.06</td>
<td>0.17</td>
<td>0.29</td>
<td>0.23</td>
<td>0.49</td>
<td></td>
</tr>
<tr>
<td>FU</td>
<td>0.07</td>
<td>0.25</td>
<td>0.27</td>
<td>0.24</td>
<td>0.52</td>
<td></td>
</tr>
<tr>
<td>Str</td>
<td>0.07</td>
<td>0.18</td>
<td>0.29</td>
<td>0.25</td>
<td>0.47</td>
<td></td>
</tr>
<tr>
<td>PA</td>
<td>0.11</td>
<td>0.13</td>
<td>0.24</td>
<td>0.26</td>
<td>0.37</td>
<td></td>
</tr>
<tr>
<td>FU</td>
<td>0.09</td>
<td>0.26</td>
<td>0.29</td>
<td>0.27</td>
<td>0.55</td>
<td></td>
</tr>
<tr>
<td>Unc</td>
<td>0.1</td>
<td>0.2</td>
<td>0.22</td>
<td>0.29</td>
<td>0.42</td>
<td></td>
</tr>
<tr>
<td>Pur</td>
<td>0.15</td>
<td>0.25</td>
<td>0.24</td>
<td>0.4</td>
<td>0.54</td>
<td></td>
</tr>
<tr>
<td>PbV</td>
<td>0.24</td>
<td>0.21</td>
<td>0.27</td>
<td>0.41</td>
<td>0.48</td>
<td></td>
</tr>
<tr>
<td>GrnT</td>
<td>0.25</td>
<td>0.17</td>
<td>0.28</td>
<td>0.42</td>
<td>0.51</td>
<td></td>
</tr>
<tr>
<td>BD</td>
<td>0.15</td>
<td>0.45</td>
<td>0.32</td>
<td>0.6</td>
<td>0.36</td>
<td></td>
</tr>
</tbody>
</table>

Table 3: Generalization experiment results.

Results

We set out to conduct a systematic comparison of several models of cross-situational word learning models across a wide variety of experimental conditions. Using a large dataset, we found best-fitting parameters for 11 models both by simultaneously fitting all of the data, and using cross-validation. The results of both fitting regimes were consistent and clear: shown in Table 1, the associative models generally achieved better fits (lower SSE and higher correlations) than the hypothesis-testing models (GrnT; Medina et al. (2011) and PbV; Trueswell et al. (2013)), which made correlated predictions (see Table 2: r = .65). The Rescorla-Wagner model, a prediction error-based associative model, performed similarly to the GrnT model, with highly correlated predictions.

Kachergis et al. (2012), and data from another nine conditions are from Kachergis et al. (2016). However, data from several conditions are contributed by Chen Yu’s lab at Indiana University, and have never before been published. Each condition consists of an ordered list of training trials consisting of 1-4 words and 2-4 objects per trial. At test, participants heard each word and were asked to select the best referent from either all m presented objects (m-alternative forced choice; AFC), or from a subset (e.g., 4AFC).

Discussion

We set out to conduct a systematic comparison of several models of cross-situational word learning models across a wide variety of experimental conditions. Using a large dataset, we found best-fitting parameters for 11 models both by simultaneously fitting all of the data, and using cross-validation. The results of both fitting regimes were consistent and clear: shown in Table 1, the associative models generally achieved better fits (lower SSE and higher correlations) than the hypothesis-testing models (GrnT; Medina et al. (2011) and PbV; Trueswell et al. (2013)), which made correlated predictions (see Table 2: r = .65). The Rescorla-Wagner model, a prediction error-based associative model, performed similarly to the GrnT model, with highly correlated predictions.
Figure 1: Human vs. model item-level accuracy using best-fitting parameters per model for all conditions (colored by condition). Note that on some items, PbV, GnT, and Pursuit all show more extreme responding (1 or 0) than people ever do.

Among the associative models, the PA model (Fazly et al., 2010) and the FU model (Kachergis et al., 2012) both fare quite well, and the sampling version of the FU model (FUs) performs slightly better, albeit with three free parameters compared to two in the PA model. The predictions of the top three models are more correlated with each other than with the human data (see Table 1): PA vs. FUs: .87, PA vs. FU: .84, FU vs. FUs: .92. But the PA model also showed a .84 correlation with the Rescorla-Wagner model, and both FU and FUs show a .87 correlation with the simpler novelty-biased associative model (N)—stronger than for the uncertainty-bias mechanism used by the FU models. Future work should aim to further explore when these models make overlapping predictions, but another approach is to find conditions where models make disparate predictions. The FUs and PA models make maximally divergent predictions in the conditions with fewer words than objects per trial (esp. 1x3, 1x3 6/18 4AFC, and 1x4), which hints at a difference in mechanism for these models. A fruitful next step would be to use optimal experiment design to generate a definitive experiment to differentiate these models.

Although we have aimed for broad coverage of models and experimental conditions, this study is far from complete in both respects. First, there are many other models that we have not yet implemented (e.g., McMurray, Horst, & Samuelson, 2012). Moreover, this dataset represents performance from

\[ r = 0.87 \] while PbV had the poorest fit, and made responses most similar to the R-W model \((r = 0.67)\). Moreover, the predictions of the R-W model were also highly correlated with the baseline co-occurrence counting model \((r = 0.999)\): such similar predictions yielded by seemingly different theories is surprising and warrants further investigation. However, some models’ particular patterns of responses suggest certain theoretical shortcomings. For example, the hybrid Pursuit model (Stevens et al., 2017) fared poorly, in part because it often outperformed humans \(M_{\text{Pur}} = 0.56, M_{\text{hum}} = 0.48\). Indeed, as shown in Figure 1, Pur, PbV, and GnT all showed a pattern of inhuman responses: for a subset of items, the models either always got them correct or always got them incorrect, while people were more stochastic.
a fairly homogeneous sample: English-speaking US college students. Future work should aim to include not only more models, but a more broadly diverse sample, both in terms of language background and age. The small number of generalization experiments run here suggest that parameters fitted to adult experimental conditions do not generalize well to developmental experiments: future research should gather a larger set of developmental data and re-fit these models with cross-validation. We invite other researchers to contribute datasets and model implementations for future, larger-scale comparisons: data and model code will be open-sourced and available on [OSF](https://osf.io).

In summary, we have shown not only how a large model comparison can help rank models from most to least able to account for human data, but also how such comparisons can identify models that mimic each other, as well as identifying experimental designs that might aid in distinguishing models. Future work should extend this comparison to cross-situational word learning models and extant datasets beyond those tested here, and design novel experiments designed to distinguish these models. Finally, assuming that cross-situational learning is an important tool in the language learner’s toolbelt, it is important to consider how these models can be extended to incorporate other information sources at the learner’s disposal, such as social cues, pragmatics, and syntactic constraints on meaning (Hirsh-Pasek, Golinkoff, & Hollich, 2000; Markman, 1990; Yurovsky & Frank, 2017). While many of these models have mechanisms for preferentially selecting and storing particular associations, most have yet to formally incorporate relevant cues from social partners—linguistic and nonlinguistic (though see Yurovsky & Frank, 2017). More broadly, we believe that structured model comparison is a critical step towards developing more complete models and, eventually, quantitative identification of the fundamental mechanisms of word learning.

### References


