Constraining the Search Space in Cross-Situational Word Learning: Different Models Make Different Predictions

Giovanni Cassani, Robert Grimm, Steven Gillis, and **Walter Daelemans**
Computational Linguistics and Psycholinguistics (CLiPS) Research Center
Department of Linguistics, University of Antwerp, 13 Prinsstraat
B-2000 Antwerpen, Belgium
{name.surname}@uantwerpen.be

Abstract

We test the predictions of different computational models of cross-situational word learning that have been proposed in the literature by comparing their behavior to that of young children and adults in the word learning task conducted by Ramscar, Dye, and Klein (2013). Our experimental results show that a Hebbian learner and a model that relies on hypothesis testing fail to account for the behavioral data obtained from both populations. Ruling out such accounts might help reducing the search space and better focus on the most relevant aspects of the problem, in order to disentangle the mechanisms used during language acquisition to map words and referents in a highly noisy environment.

Keywords: cross-situational learning; word learning; computational modeling; language acquisition

Ever since the *gavagai* example provided by Quine (1960) to describe the huge amount of referential uncertainty that any language learner has to face while inducing word-object mappings, researchers took an interest in which mechanisms can be exploited to solve this crucial task. In the last twenty years, computational modeling has proven extremely useful in exploring what information encoded in the input children receive might allow them to correctly map referents and words, and which learning mechanisms might best exploit the relevant information (Frank, Goodman, & Tenenbaum, 2009).

Cross-situational learning posits that children keep track of co-occurrences of referents in the world and words uttered to them in several situations to establish unambiguous mappings: while the single situation might be ambiguous, the co-occurrences of words and referents across many different situations help the learner figure out the correct mappings. Starting from the work of Siskind (1996), many different learning mechanisms that exploit this basic principle have been proposed that show comparable performances to many behavioral data from both children (Ramscar, Dye, & Klein, 2013; Smith & Yu, 2008; Suanda, Mugwanya, & Namy, 2014) and adults (Dautriche & Chemla, 2014; Fazly, Alishahi, & Stevenson, 2010; Medina, Snedeker, Trueswell, & Gleitman, 2011; Trueswell, Medina, Hafri, & Gleitman, 2013; Yu & Smith, 2007; Yurovsky & Frank, 2015; Yurovsky, Yu, & Smith, 2013), using both corpus studies and laboratory experiments, covering many different conditions. Differences and similarities across models have been explored, with the main goal of showing how apparently different proposals can yield comparable results and make similar predictions when certain components of the learning algorithm are modified (Yu & Smith, 2012; Kachergis, Yu, & Shiffrin, 2016).

In this paper, we compare behavioral evidence to four different models that exploit cross-situational regularities to infer word-referent mappings from the data, to analyze what predictions each model makes and whether they fit with what children and adults do when asked to map a referent to a word. Our aim is to provide evidence about which learning mechanisms proposed in the literature can explain behavioral evidence and which cannot, in order to constrain the search space of possible models to the learning strategies that exploit cross-situational information in the same way humans do. Carefully controlled laboratory settings in which specific features of the word learning task are manipulated can help to achieve this goal, by isolating the information from the input that makes learning possible or impossible and providing valuable data to test computational simulations in a variety of situations (Ramscar, Dye, & Klein, 2013; Kachergis et al., 2016).

While many learning mechanisms can mirror certain behavioral patterns (Yu & Smith, 2012), some may not be able to learn the correct word-referent mappings in specific, controlled paradigms in which subjects do learn such mappings robustly. Identifying these situations and showing why certain mechanisms fail to account for successful learning will help the researchers to constrain the hypothesis space and discard mechanisms that make incorrect predictions.

Dataset

In order to evaluate the predictions of different models of cross-situational learning we make use of the evidence presented in Ramscar, Dye, and Klein (2013). The experiment they reported was conducted with a group of children (mean age 28 months) and two groups of adults, undergraduates and developmental psychologists.

The setting included three objects, [ObjA, ObjB, ObjC], and three labels, $\{Dax, Wug, Pid\}$; during 18 learning trials, each subject saw two objects and then heard one label. Of the three objects, ObjA and ObjC were presented 9 times, never together; ObjB, however, was present in all trials, occurring half of the times with ObjA and half of the times with ObjC. Crucially, ObjA was always presented together with the same label, e.g. *Dax*, and ObjC was always presented with the same label, e.g. *Pid*. Consequently, ObjB occurred half of the times with the label *Dax* and half of the times with the label *Pid*. The third label, *Wug*, never occurred during training.

During testing, the subjects heard one of the three labels

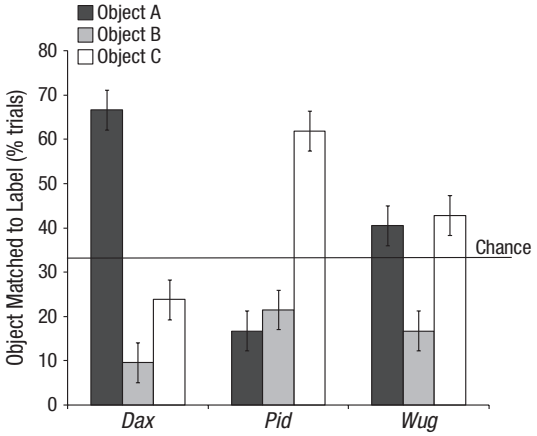

Figure 1: Children learning patterns on the word-learning experiment from Ramscar, Dye, and Klein (2013). The plot is taken from the original paper (Figure 2a).

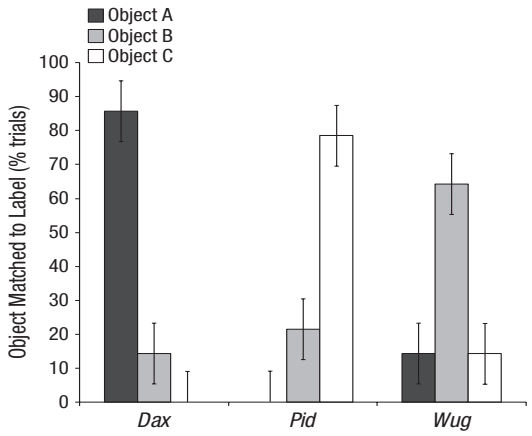

Figure 2: Undergraduates learning patterns on the word-learning experiment from Ramscar, Dye, and Klein (2013). The plot is taken from the original paper (Figure 3a).

and were asked to point to the object to which they thought the label referred to. Two labels occurred during training, one did not - words and objects were counterbalanced and learning trials were randomized across participants.

Behavioral Evidence

Results of the experiment are provided in Figure 1 for children and Figure 2 for undergraduates - the plots show the case in which *Dax* was always presented with ObjA, *Pid* with ObjC, and *Wug* was only showed during testing.

Both groups mapped ObjA and ObjC to the labels that only occurred with each of them. Interestingly, however, undergraduates showed a mutual exclusivity bias and mapped ObjB to *Wug*, which was not presented during training; on the contrary, children picked ObjA and ObjC at comparable rates as referent for the new label. The developmental psychologists were asked to predict the behavior of children but ended up predicting that of undergraduates. The authors of the study conclude that children are more sensitive to the informativity of cues than to logical principles, which on the contrary play a role in adults.

Feature-Label-Order Effects In this experiment, and in many others that address cross-situational word learning, objects are presented before their labels are uttered. Far from being irrelevant to the task, evidence from Ramscar, Yarlett, Dye, Denny, and Thorpe (2010) shows that different learning outcomes arise in behavioral experiments where this order is manipulated. This difference is unfortunately not always considered in cross-situational learning studies: as a consequence, certain models are defined as mapping referents to words and others do the opposite. Moreover, the behavioral data we use were obtained using a paradigm in which the subjects first saw an object and then heard a label. Thus, considering the experimental paradigm and the importance of the order of presentation for learning to take place, when we evaluate a model that was designed to map words to referents, we switch the two layers and make it learn the opposite mapping.

Models of Cross-situational Learning

We compare simple, basic implementations[1] of four different learning mechanisms to highlight what predictions are made by each of them, and whether they match behavioral evidence. We introduce each model separately and briefly discuss its main features; for more detailed explanations, we refer to the cited publications.

Hebbian Learner This model implements the law of contiguity (Warren, 1921), according to which the association between two items becomes stronger when they consistently occur together in the environment. It is usually implemented as a neural network with no hidden layer that incrementally establishes associations between an input and an output layer (Hebb, 1949). An input-to-output association is strengthened by a constant quantity whenever the two co-occur within a learning trial. Associations from inputs that occur in a learning trial and outputs that do not are left unchanged, as are associations from absent inputs to all output units.

The way associations are updated is summarized in equation (1), where t represents a learning trial, c_i indicates an input item, or cue, o_j indicates an output, or outcome, and ΔV_{ij} indicates the value of the update from c_i to o_j after experiencing the learning trial t:

$$\Delta V_{ij} = \begin{cases} k & \text{if } c_i \in t \text{ and } o_j \in t \\ 0 & \text{else} \end{cases} \quad (1)$$

[1]The code of our own re-implementations of each model is available at `https://github.com/GiovanniCassani/cross_situational_learning`, commit n. 2a9dbaa

ΔV_{ij} is then added to the current association from c_i to o_j; k is a strictly positive constant which only affects the absolute value of the associations but not the relations among them, thus changing its value does not affect the learning outcome. This model was showed to successfully model behavioral data in the study by Yu and Smith (2012) and for this reason it is evaluated here. However, the risk exists that every input becomes associated with every output, making it impossible to learn unambiguous input-output mappings (Dawson, 2008).

Naïve Discriminative Learning (NDL) In this model, input-output associations are updated according to the Rescorla-Wagner equations (Rescorla & Wagner, 1972), developed in the context of animal learning and conditioning. This model is often referred to as Naïve Discriminative Learning (NDL, (Baayen, Hendrix, & Ramscar, 2013)) and its relevance to language has been established in different aspects of language learning and processing (Baayen, Milin, Durdević, Hendrix, & Marelli, 2011; Baayen, Shaoul, Willits, & Ramscar, 2015; Ramscar, Dye, & McCauley, 2013; Ramscar, Hendrix, Shaoul, Milin, & Baayen, 2014).

Its architecture closely resembles the Hebbian learner, as it is a neural network with no hidden layer that incrementally establishes associations between cues and outcomes, where the first constitute the input layer and the latter the output nodes. As for the Hebbian learner, when a cue co-occurs with an outcome, the association between them becomes stronger; moreover, associations from absent cues (in a learning trial) to all outcomes are left unchanged. However, in the NDL model, associations from present cues to absent outcomes are weakened, and can eventually become negative. The model is naïve because every outcome is updated independently of all other outcomes.

The update in associations is summarized in equation (2), where t is a learning trial, ΔV_{ij} is the change in association involving a cue c_i and an outcome o_j.

$$\Delta V_{ij} = \begin{cases} \alpha_i\beta_1(\lambda - \sum_{c\in t} V_c) & \text{if } c_i \in t \text{ and } o_j \in t \\ \alpha_i\beta_2(0 - \sum_{c\in t} V_c) & \text{if } c_i \in t \text{ and } o_j \notin t \\ 0 & \text{if } c_i \notin t \end{cases} \quad (2)$$

α_i is a parameter modifying the salience of an input unit, or cue: while a different value can be set for each cue, this parameter is usually kept constant to remain agnostic with respect to cue importance. β_1 and β_2 specify the importance of positive and negative evidence respectively. These two parameters can again take different values but are usually set to the same quantity to reduce the initial assumptions. λ is the maximum amount of association that each outcome can receive from all inputs and operates as a simple linear scaling factor (Evert & Arppe, 2015). Finally, $\sum_{c\in t} V_c$ is the total association supported by the cues present in the current learning trial: this evidence is used to predict the outcome, and the prediction error is used to update cue-outcome associations. ΔV_{ij} is added to the current association value of cue c_i for each outcome o_j encountered up to trial t. The same happens for all $c_i \in t$. For the reported simulations we selected standard parameter values that allow to make minimal assumptions, setting all $\alpha s = 0.2$; $\beta_1 = \beta_2 = 0.1$; $\lambda = 1$.

Probabilistic learner In its original formulation (Fazly et al., 2010), this model computes a posterior probability distribution over referents for each word, updating the probability mass allocated to each referent in the light of new evidence. A referent r that seldom occurs with a word w but often occurs with many other words will get a small probability for w, while a referent r' that often occurs with word w and rarely with others will have a high probability of being the correct referent for w. The model incrementally updates associations between words and referents and uses them to compute the conditional probability of a referent given a word for all the referents that occurred with the word up to the present learning trial.

More generally, this model can be thought of as computing a posterior distribution over all possible outcomes for each cue. Associations between cues and outcomes are computed as specified in equations (3-5), where t is a learning trial, o is an outcome from the set of outcomes in the learning trial, O_t, c is a cue, from the set of cues in the learning trial, C_t, paired with O_t, and C is the set of cues encountered up to t:

$$a(c|o, O_t, C_t) = \frac{p_{t-1}(o|c)}{\sum_{c'\in C_t} p_{t-1}(o|c')} \quad (3)$$

$$assoc_t(c,o) = assoc_{t-1}(c,o) + a(c|o, O_t, C_t) \quad (4)$$

$$p_t(o|c) = \frac{assoc_t(c,o) + \lambda}{\sum_{o'\in O} assoc_t(c,o') + \beta\cdot\lambda} \quad (5)$$

This model has 3 free parameters. λ is a small smoothing factor; β is the upper bound on the expected lexicon; $p_{t=0}(o|c)$ is the initial value of the probability of an outcome given a cue, before they are encountered in a learning trial. In the simulations reported by Fazly et al. (2010), $\beta = 8.500$, $\lambda = 10^{-5}$, and $p_{t=0}(o|c) = 1/8{,}500$. We kept the same value for λ, set $\beta = 10^4$ and $p_{t=0}(o|c) = 10^{-4}$.

Equation (3) computes the update in association between a cue and an outcome from the current learning trial: this update is proportional to $p(o|c)$ at the previous learning trial and depends on the number of cues in the current trial: more cues cause a lower change, due to higher noise and uncertainty in the current trial. The update computed in (3) is added to the corresponding cue-outcome association, as specified in (4). Associations are not exploited directly but rather used to update a probability distribution over outcomes for each cue. More evidence makes the learner allocate a higher posterior probability to a specific outcome. In (5), the denominator acts as a scaling factor that implements within-trial competition: if a cue c is already associated to one of the previously encountered outcomes, the probability that c maps to another outcome does not receive strong support.

In the original formulation, words were cues and referents were outcomes; however, considering what we discussed in the section about Feature-Label-Order effects (Ramscar et al., 2010), we flipped the encoding so that this algorithm learns a probability distribution for words over referents, coding words as outcomes and referents as cues[2].

Hypothesis-Testing Model (HTM) The HTM model selects, stores and updates a single hypothesis for each learning trial. Initially, it randomly picks a word-referent mapping from the possible ones in the learning trial. When an already encountered word is presented in a subsequent trial, the model looks in memory to retrieve the hypothesized referent for the word and may retrieve it or not. If it does, the hypothesis is strengthened when confirming evidence is found in the current trial and discarded otherwise, in which case a new referent is hypothesized at random for the word being considered. If no hypothesis is recalled, a new referent is hypothesized at random for the word being considered and the old one fades away. As is specified in Medina et al. (2011) and Trueswell et al. (2013), the model depends on one main parameter, α, which models the probability that a formed hypothesis is retrieved from memory. However, Trueswell et al. (2013) argue that the value of this parameter changes when a hypothesis is recalled: the next time the label appears, the hypothesized referent should be retrieved with a higher probability if it was already retrieved. Unfortunately, however, no function was specified to model the change of α after successful retrievals. Therefore, we set the initial and second values for α at 0.6 and 0.81, following the third experiment in Trueswell et al. (2013), were this model was shown to fit behavioral results. Accordingly, the first time a hypothesis can be retrieved with probability equal to 0.6; if it gets confirmed, the next time it will be retrieved with probability equal to 0.81. Since we have many more trials, we set further values, 0.9, 0.95, and 0.99, to model the probability that a hypothesis is retrieved after the third, fourth or fifth time it was retrieved and confirmed. After the fifth time α does not change anymore: we stopped at 0.99 to exclude certainty of recall.

Computational Simulations

In order to closely mimic the learning task that was faced by children in the study by Ramscar, Dye, and Klein (2013), we implemented incremental learners: the connection between a referent and a word is only updated when both have been encountered in a learning trial. This is crucially different from the simulations implemented by Ramscar, Dye, and Klein (2013), where the equilibrium equations (Danks, 2003) of the NDL model (Baayen et al., 2011) were used. In this case, the end state of the model is computed when no more learning trials are available. Equilibrium equations have the advantage of not depending on any free parameter, but all cues and all outcomes are simultaneously available to the learner, differently from the task faced by the subjects. There was no way they could expect a third label to be presented during testing and thus update connections from objects to that label during training.

Here, we focus on the situation where children and undergradutes showed consistent behaviors, i.e. in retrieving an object when presented with a label they encountered during training. If a model fails to account for this aspect of the data, it can be hardly justified as a model of human cross-situational learning, during acquisition as well as in adulthood. Accordingly, we train each simulation using the input presented to the subjects and evaluate the final state of learning. However, since different models learn different things (associations, probabilities, hypotheses), it is hard to directly compare them. We do not assume any linking mechanism that converts internal representations to behavior; we simply look at the learned representations and evaluate whether unambiguous mappings were learned, that could allow subjects to retrieve an object, consistently with learning displayed by human subjects.

In each learning trial, simulated learners were given a set of objects and a word: beside ObjA, ObjB, and ObjC, the set of cues also contained other cues that account for the whole experimental context,[3] for consistency with the original simulation in Ramscar, Dye, and Klein (2013). Table 1 summarizes the input to the computational models.

Table 1: Training trials, as described in Ramscar, Dye, and Klein (2013)

Cues	Outcomes	Freq
ObjA_ObjB_Context1_ExptContext	Dax	9
ObjB_ObjC_Context2_ExptContext	Pid	9

For all models, we ran 200 simulations randomizing the order of presentation of the learning trials: since no model depends on initial random values, the order of the trials is the only potential source of bias. We report referent-word associations at the end of training for the four models in Table 2[4].

Successful learning happens when, for each label, the value corresponding to an object is consistently higher than the values of the other two objects, given that the test procedure consisted of presenting a label and asking for the matching object. In this setting, the Hebbian learner would choose randomly and is not learning much, since, in both the *Dax* and *Pid* columns, two objects have the same association to each label. On the contrary, the NDL model would retrieve the correct object given the two words provided during training, since the ObjA-*Dax* and ObjC-*Pid* associations are higher than any other. Another interesting feature is that it learns that ObjA does not come with the label *Pid*, forming a neg-

[2]Personal communication with one of the authors confirmed that the learning mechanism is not altered by switching the mapping.

[3]This was not the case for the HTM model, in which only ObjA, ObjB, and ObjC were provided as input.

[4]For explanatory purposes we will focus on the three objects, leaving the other cues out.

Table 2: Referent-word associations after 18 training trials (200 simulated learners). For the Probabilistic Learner, conditional probabilities of label given object are showed; for the Hypothesis Testing Model, the proportion of learners that selected each hypothesis is showed.

Model	Cue	Dax	Pid
Hebbian Learner	ObjA	9	.
	ObjB	9	9
	ObjC	.	9
NDL	ObjA	.134 $\pm$.001	-.021 $\pm$.005
	ObjB	.113 $\pm$.005	.113 $\pm$.005
	ObjC	-.021 $\pm$.005	.134 $\pm$.001
Probabilistic Learner	ObjA	.967 $\pm$.003	.
	ObjB	.484 $\pm$.085	.485 $\pm$.085
	ObjC	.	.967 $\pm$.003
HTM	ObjA	.465	.
	ObjB	.535	.53
	ObjC	.	.47

ative association. The Probabilistic Learner makes similar predictions to the NDL model, except for the negative associations. Finally, the HTM performs close to chance, with as many simulated learners mapping *Dax* to ObjA as to ObjB, and *Pid* to ObjB and ObjC, again showing no sign of learning, inconsistently with the behavioral evidence we considered.

Discussion

Our results show that some of the proposed learning mechanisms fail to account for the behavioral data obtained by Ramscar, Dye, and Klein (2013), for both children and adults: specifically, a Hebbian learner (Hebb, 1949) and the HTM (Trueswell et al., 2013) fail to learn robust object-label mappings. Two other models, the Probabilistic Learner (Fazly et al., 2010) and the NDL model (Baayen et al., 2011), show remarkably similar patterns to the behavioral data from both children and adults. The behavioral evidence also makes it clear that it is not necessary for successful cross-situational learning that true word-referent associations are more frequent than spurious associations. As a matter of fact, in the dataset each word-referent pair occurs with the same frequency, defying the very notion of a spurious pairing: ObjA could be paired to *Dax* just as ObjB could, if we only consider frequency of co-occurrence of objects and words. Nonetheless, humans learned consistent mappings, suggesting that simply tracking co-occurrence frequencies is a poor candidate mechanism to explain cross-situational word learning.

As is often the case in attempts to compare models, many decisions need to be taken and different choices can result in different outcomes. For example, we did not equip the HTM with a mutual exclusivity bias, mainly because it is not specified in the paper where the model was proposed and also because we wanted to evaluate basic versions of each model to focus on the proposed learning mechanisms rather than specific features. However, even with such a bias, the HTM would fail to match the behavioral data. Consider the situation in which the model first sees a *Dax* trial and it randomly picks ObjB as a referent. When a *Pid* trial is presented, the learner searches in memory, finds a *Dax*-ObjB hypothesis, decides that *Pid*-ObjB is not legitimate, and maps *Pid* to ObjC. If the HTM starts with a wrong mapping for *Dax*, it will only find the correct mapping for *Pid*, but will keep failing at relating ObjA to *Dax*. The problem lies in the single hypothesis assumption, not in the absence of the mutual exclusivity bias. In order to account for this behavioral evidence, a model should hold in memory the two possible hypotheses. Only then could it appreciate the fact that ObjB occurs with both labels while ObjA and ObjC consistently occur with one. The same problem of failing to appreciate the different background rates of the three objects affects the Hebbian learner, but results from an entirely different architecture, since it only focuses on co-occurrences to update associations. However, the behavioral evidence suggests that subjects do assign importance to missing co-occurrences too, and our simulations show that successful learning is only possible when a model is sensitive to both positive and negative co-occurrences. Taken together, the failures of the HTM and the Hebbian learner point to the importance of storing multiple mappings and being sensitive to both things that co-occur and things that fail to co-occur in the environment (Ramscar, Dye, & McCauley, 2013).

Unlike Trueswell et al. (2013) and Dautriche and Chemla (2014), we only evaluated the end-state of learning and did not consider trial-to-trial patterns, due to the behavioral data we used for comparison. This analysis would have certainly been useful because it allows to follow the learning trajectory. However, if a model does not account for the end state of learning it can hardly explain the mid-states, while a model that fits the final picture might have done so in different ways than the subjects. Thus, the reported evidence appears to be strong enough to make a case against the psychological plausibility of a model, while more evidence is needed about models that fit the behavioral data.

Finally, we did not evaluate any specification of which mechanism can make use of the associations learned during training to actually decide which object to retrieve when presented with a new label. While this is an interesting component of the paradigm in Ramscar, Dye, and Klein (2013) and it is crucial to investigate how learning mechanisms differ between young children and adults, we provided evidence that some learning mechanisms fail to account for behavioral data from both groups even when the much simpler condition of retrieving a referent when presented with a known word is considered. Further analyses are required to identify those mechanisms that can both i) form the correct associations during training and ii) use such associations to retrieve a known referent for an unknown word, in the same way children and adults do, to highlight where their learning mechanisms differ and where they are comparable.

Conclusion

The evidence we provided in this paper complements the study by Yu and Smith (2012) by showing that not every learning mechanism can be instantiated in an algorithm that accounts for behavioral data in cross-situational word learning. A single-hypothesis learning strategy (Medina et al., 2011; Trueswell et al., 2013) and an associative model that only relies on Hebbian learning (Hebb, 1949) fail to fit behavioral data. The jury is still out about the Probabilistic Learner (Fazly et al., 2010) and the Naive Discriminative Learner (NDL, (Baayen et al., 2011)): both models fit the results by Ramscar, Dye, and Klein (2013), but they behave differently, prompting for further research on which mechanisms underpin cross-situational learning in humans.

Acknowledgments

We are grateful to Michael Ramscar and Konstantin Sering for the discussion over the NDL model; to Aida Nematzadeh for her help in better understanding the Probabilistic Learner; to Chen Yu for sharing his latest research with us.
This research was supported by a BOF/TOP grant (ID 29072) of the Research Council of the University of Antwerp.

References

Baayen, R. H., Hendrix, P., & Ramscar, M. (2013). Sidestepping the combinatorial explosion: An explanation of n-gram frequency effects based on naïve discriminative learning. *Lang Speech*, *56*(3), 329-347.

Baayen, R. H., Milin, P., Durdević, D. F., Hendrix, P., & Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naïve discriminative learning. *Psychol Rev*, *118*(3), 438-481.

Baayen, R. H., Shaoul, C., Willits, J., & Ramscar, M. (2015). Comprehension without segmentation: a proof of concept with naïve discriminative learning. *Lang Cogn Neurosci*, 1-23.

Danks, D. (2003). Equilibria of the Rescorla–Wagner model. *J Math Psychol*, *47*(2), 109-121.

Dautriche, I., & Chemla, E. (2014). Cross-situational word learning in the right situations. *J Exp Psychol Learn Mem Cogn*, *40*(3), 892.

Dawson, M. R. W. (2008). Connectionism and classical conditioning. *Comp Cogn Behav Rev*, *3. Monograph*, 1-115.

Evert, S., & Arppe, A. (2015). Some theoretical and experimental observations on naïve discriminative learning. In *Proceedings of the 6th Conference on Quantitative Investigations in Theoretical Linguistics.*

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

Frank, M. C., Goodman, N. D., & Tenenbaum, J. B. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychol Sci*, *20*(5), 578-585.

Hebb, D. O. (1949). *The organization of behavior*. New York, NY: John Wiley and Sons.

Kachergis, G., Yu, C., & Shiffrin, R. M. (2016). A bootstrapping model of frequency and context effects in word learning. *Cognitive Science*.

Medina, T. N., Snedeker, J., Trueswell, J. C., & Gleitman, L. R. (2011). How words can and cannot be learned by observation. *Proc. Natl. Acad. Sci.*, *108*(22), 9014-9019.

Quine, W. V. O. (1960). *Word and object*. Cambridge, MA: MIT Press.

Ramscar, M., Dye, M., & Klein, J. (2013). Children value informativity over logic in word learning. *Psychol Sci*, *24*(6), 1017-1023.

Ramscar, M., Dye, M., & McCauley, S. M. (2013). Error and expectation in language learning: The curious absence of mouses in adult speech. *Language*, *89*(4), 760-793.

Ramscar, M., Hendrix, P., Shaoul, C., Milin, P., & Baayen, R. H. (2014). The myth of cognitive decline: non-linear dynamics of lifelong learning. *Top Cogn Sci*, *6*(1), 5-42.

Ramscar, M., Yarlett, D., Dye, M., Denny, K., & Thorpe, K. (2010). The effects of Feature-Label-Order and their implications for symbolic learning. *Cognitive Science*, *34*, 909-957.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: current research and theory* (p. 497). New York, NY: Appleton-Century-Crofts.

Siskind, J. M. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, *61*(1), 39-91.

Smith, L. B., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross- situational statistics. *Cognition*, *106*(1), 1558-1568.

Suanda, S. H., Mugwanya, N., & Namy, L. L. (2014). Cross-situational statistical word learning in young children. *J Exp Child Psychol*, *126*(1), 395–411.

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

Warren, H. C. (1921). *A history of the association psychology*. New York, NY: Charles Scribner's Sons.

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

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

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

Yurovsky, D., Yu, C., & Smith, L. B. (2013). Competitive processes in cross-situational word learning. *Cognitive Science*, *37*(5), 891-921.