A Psychologically Motivated Model of Word Learning

Jon Stevens (jonsteve@ling.upenn.edu), Charles Yang, John Trueswell, Lila Gleitman
University of Pennsylvania

We present an on-line computational model that aims to bridge the gap between formal and psychological studies of word learning. Like the model of Fazly et al. (2010), our model incrementally processes input data rather than using iterative optimization to derive the lexicon over the entire corpus of learning instances (Yu & Ballard 2007, Frank et al. 2009). Unlike Fazly et al. (2010), our model incorporates recent experimental findings (Medina et al. 2010, 2011) which suggest a simpler learning mechanism than that assumed by most recent computational studies beginning with Siskind (1996). Our approach models word learning with a general class of mechanisms in which linguistic hypotheses (word-to-meaning mappings) are probabilistically assessed and evaluated against input data, similar to the variational model of parameter setting (Yang 2002). Following experimental results, our learner only considers one semantic hypothesis at a time for each word (partial cross-situational knowledge) rather than keeping track of all possibilities (full cross-situational knowledge). We show that this simpler mechanism can produce competitive results when tested on annotated child-directed English input.

Motivation It is usually assumed that children learn the meanings of their first words by storing a large number of situations during which a word is uttered in memory and then finding the commonalities across these experiences. Two recent studies have challenged this idea. Instead, they suggest that learners consider one candidate meaning at a time, and converge to the target by making and testing hypotheses of word meaning against sequences of learning instances.

Medina et al. (2010) used a procedure in which subjects heard nonsense words accompanying objects on a screen and were asked, for each word, to click on the object they thought it referred to (henceforth "clicking experiment"). Each word occurred with five different sets of objects, with an object representing the target meaning always present. Subjects could receive a highly informative (HI) instance (with only two objects to choose from rather than five) either as the first (HI first), third (HI middle), or last (HI last) instance for each word. If word learners were really accumulating full contextual data with each instance, one would expect performance on later instances to be significantly better than performance on earlier instances. However, subjects perform at chance level (50%) for all HI instances (see the left side of Fig. 1). Having a HI instance in the middle of the experiment was no better than having it at the beginning. Using the human simulation paradigm (Gillette et al. 1999, Gleitman et al. 2005), Medina et al. (2011) studied the trajectory of word learning over sequentially presented learning instances. Subjects were shown video vignettes along with nonsense words and asked to give verbal guesses as to the meaning of the nonsense word after each video and then a final conjecture for each word (henceforth "guessing experiment"). Here the informativeness of a learning instance was assessed according to how easy it was to determine the correct meaning after viewing the vignette in isolation (as determined in a separate experiment). As in Medina et al. (2010), HI instances were helpful only for the first learning instance, as they were of use only insofar as they lower the probability of an incorrect initial guess, making the learner less likely to waste time testing erroneous hypotheses. In addition, Medina et al. (2011) examined individual learning trajectories, finding that people appear to evaluate just the previously hypothesized meaning against the current learning instance: for example, if the learner makes the wrong guess for an initial HI instance, the selection of the target meaning for the second learning instance is at chance, indicating little or no memory of the correct alternative referent that could have been inferred from the initial HI instance (see Fig. 4).

The present study demonstrates the effectiveness of this type of learning on real child-directed speech data, showing that partial cross-situational knowledge is sufficient to learn from a small number of utterance-situation pairs.

The Present Approach The learner's sequence of hypothesis testing against data strongly resembles well known models of Reinforcement Learning that have roots in both mathematical psychology (e.g., Bush & Mosteller 1951) and machine learning (Narendra & Thathacher 1989, Barto & Sutton 1998). In these models, the learner assigns a probabilistic distribution to a set of hypotheses/choices. At every instance of learning, a hypothesis is selected according to its probability and evaluated against the input data: confirmation is rewarded by increasing the probability of the selected hypothesis and disconfirmation is penalized by decreasing the probability. This type of learning model has been adapted for syntactic

parameter setting in the variational model of language acquisition (Yang 2002). In both cases, the learner determines the probabilistic distribution over a set of hypotheses through exposure to linguistic input.

The learning model is illustrated in Fig 3. Each word is associated with a probabilistic vector over the set of possible meanings, i.e., objects in the environment when the utterance containing the word is produced. (The utterance is a single word in our replication of the Medina et al. (2010, 2011) studies, and is a sentence in our simulations using child-directed English data.) The probabilistic vectors are used to select hypotheses of word meanings when new utterances are encountered. If the object to which the chosen meaning refers is present in the environment, the meaning is rewarded. Otherwise it is penalized. Probabilities are renormalized after each instance of learning. In other words, if one meaning is rewarded, then the probabilities of all other meanings are decreased. These vectors are then transformed using a threshold into a discrete lexicon of word-to-meaning mappings.

Since the human simulation results show that an initially correct hypothesis of word meaning greatly enhances the accuracy of learning, we have adapted the so-called pursuit algorithm (Barto & Sutton 1998), a variant of Reinforcement Learning (Bush & Mosteller 1951, Yang 2002), to capture this effect. Pursuit learning gives an advantage to the most probable hypothesis—one which may not be correct—so that its probability of being selected for evaluation is higher than what its probability warrants.

This model is able to reproduce key aspects of the human experimental results described above. For the clicking experiment, the informativeness is trivially encoded in the number of objects there are to choose from. In the guessing experiment, we used the non-cross-situational (isolation) experimental results as a proxy for how salient a particular meaning is within a vignette. For example, if 7 out of 12 people guessed the meaning 'bag' after seeing a 'bag' vignette in isolation, then the initial probability for the "mipen"-"bag" mapping was set to 7/12; HI instances are defined as those where the correct meaning has an initial probability of ½ or higher. An initial highly informative instance makes a correct guess more likely, which will receive continued confirmation as new data is processed. Thus, in both experiments the HI first group has an overall advantage. This is shown for the clicking experiment in Fig. 1 and for the guessing experiment in Fig. 2. For the guessing experiment with a HI first learning instance, if the learner makes an initial wrong guess, the pursuit algorithm will increase its probability of being chosen for the second instance leading to a failure. As a result, one of the other hypotheses, including the correct one, will be chosen with equal probability, again mirroring the performance of human subjects. We see in Fig. 4 that the average accuracy for those who get it wrong the first time is always below 10%.

Evaluation on Child-directed Speech Does the model scale up in actual settings of language acquisition? To this end, we evaluated the model on data from the Rollins corpus from the CHILDES database and compared performance to that of the more complex Bayesian model of Frank et al. (2009) as well as the incremental Machine Translation-based model of Fazly et al. (2010). Unlike the experimental simulations, the Rollins data does not directly encode the informativeness of an utterance. However, infants use social cues like eye gaze and gestures to direct their attention to intended referents (Baldwin 1991, Bloom 2000, Nappa et al., 2009). Furthermore, infants are sensitive to prosodic information, especially the exaggerated peaks and valleys of child-directed speech (Soderstrom, Seidl, Nelson & Jusczyk, 2003). We follow Yu & Ballard (2007) in making use of both gestural and prosodic cues by assigning greater weight to mappings between prosodically salient words (stressed and at phrasal boundaries) and gesturally indicated objects. The child-directed speech with accompanying video was coded for saliency by hand.

Performance is evaluated by the usual metric of precision and recall and the F-score computed from these measures (Table 1). Since the output lexicon differs between simulations, precision and recall numbers for our model are the average of 100 simulations. The present model is competitive against Frank et al. (2009), which is considerably more complex and cannot run incrementally, and slightly outperforms the more plausible Fazly et al. model.

Conclusion These results show that insights from the behavioral studies of language learning can simplify the computational problem of early word learning. A psychologically motivated model that considers only one hypothesis at a time is competitive to or better than models that give the learner full knowledge of every situation.

Figures and tables

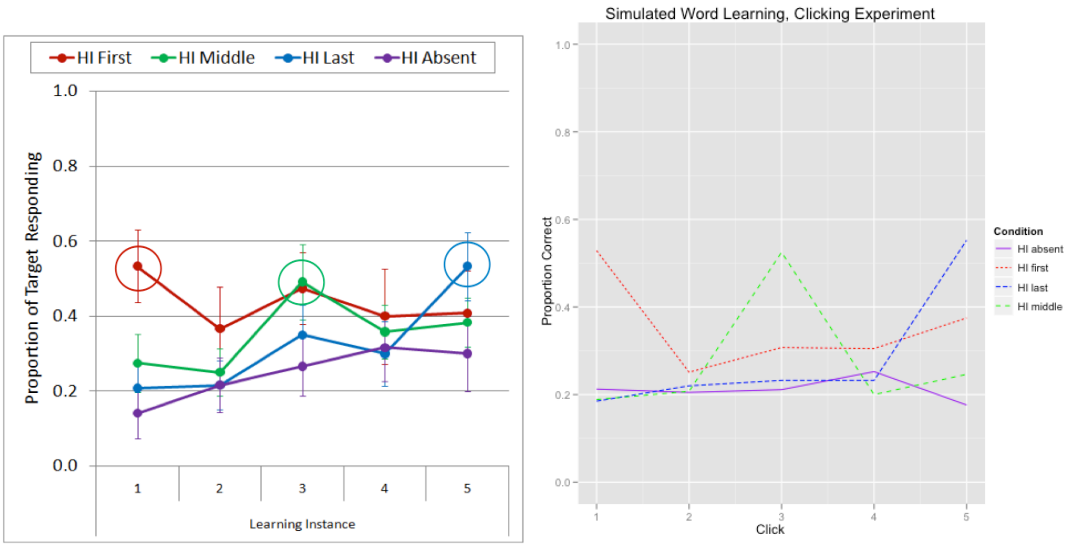

Figure 1: Human simulated learning (left) vs. machine simulated learning (right), clicking experiment

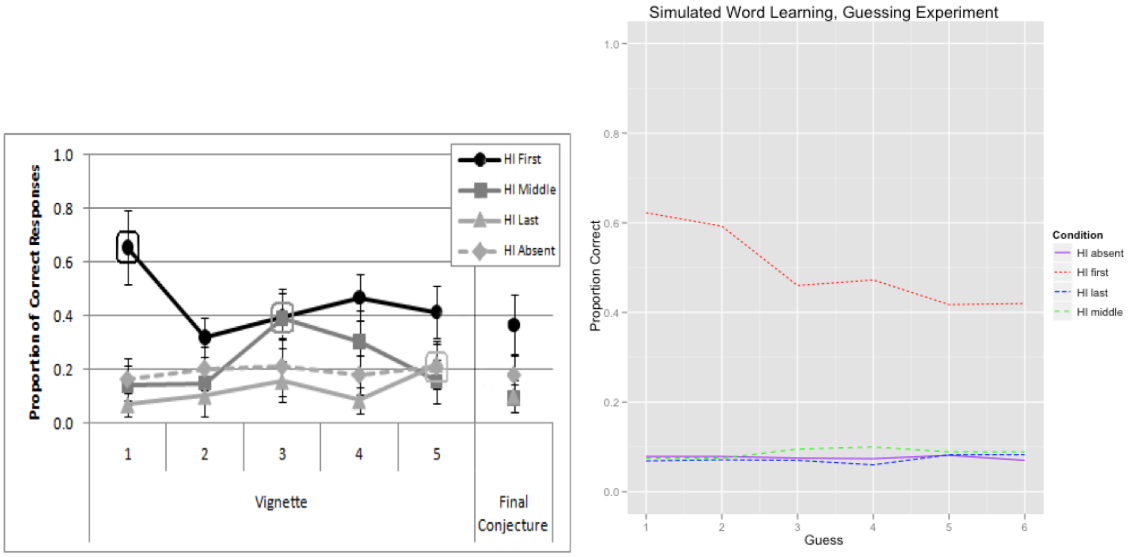

Figure 2: HSL vs. MSL, guessing experiment

For each word in an utterance, accompanied by a set of visible word meanings:
  If the word is novel:
    Initialize a probability vector for that word consisting of equal probabilities for all visible meanings.
    Initialize a vector of value estimates for that word with all values set to zero.
    Choose an initial hypothesis H from the set of visible meanings.
  Else:
    Introduce any new meanings into the probability vector with a non-trivial probability of being chosen.
    Use the probability vector to choose a hypothesis H.
  Evaluate H and update the word's probabilities and estimated values using the pursuit learning algorithm.
  Build a lexicon of all word-to-meaning mappings with an estimated value greater than a given threshold.

Figure 3: A psychologically motivated word learning algorithm

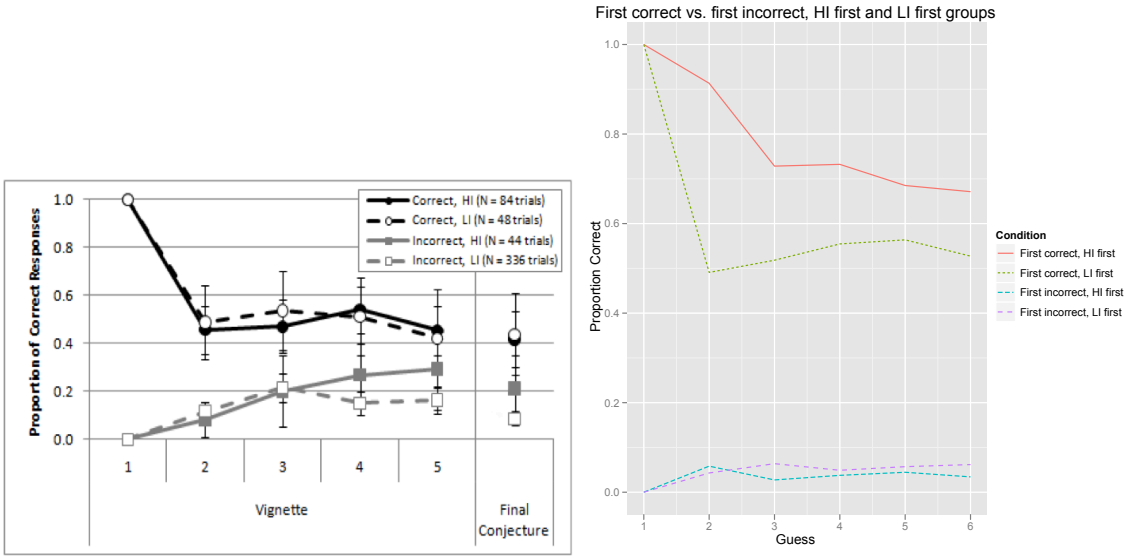

Figure 4: Accuracy by whether the first guess was correct (HI first and LI first), guessing experiment (Human on left, machine on right)

	Precision	**Recall**	**F-score**
Bayesian, no cues	0.36	0.29	0.32
Bayesian, prosody & gesture	0.72	0.38	0.52
MT, no cues	0.23	0.09	0.13
MT, prosody & gesture	0.29	0.55	0.35
Current, no cues	0.21	0.39	0.28
Current prosody & gesture	0.30	0.62	0.40

Table 1: Current model compared with Bayesian (Frank et al. 2009) and MT (Fazly et al. 2010) models (Precision and recall for Current are averages over 100 simulations)

References

Baldwin, D. (1991). Infants' contribution to the achievement of joint reference. *Child Development, 62(5),* 874-890.

Barton, A. & R. Sutton. (1998). *Reinforcement Learning.* Cambridge, MA: MIT Press.

Bloom, P. (2000). How children learn the meanings of words. Cambridge, MA: MIT Press.

Bush, R. & F. Mosteller (1951). A mathematical model for simple learning. *Psychological Review*, *58*, 313-323.

Frank, M., Goodman N. & Tenenbaum J. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science 20:5*, 578-585.

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

Gleitman, L.R., Cassidy, K., Papafragou, A., Nappa, R., & Trueswell, J.C. (2005). *Hard words*. Journal of Language Learning and Development, *1(1)*, 23-64.

Medina, T.N., Hafri, A., Trueswell, J.C. & Gleitman, L.R. (2010). Propose but verify: Fast-mapping meets cross-situational word learning. Paper presented at the BUCLD. Boston, MA.

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 Science, 108(22),* 9014-9019

Nappa, R., Wessel, A., McEldoon, K.L., Gleitman, L.R. & Trueswell, J.C. (2009). Use of speaker's gaze and syntax in verb learning. *Language Learning and Development*, *5*, 203-234

Narendra, K., & Thathachar, M. (1989). *Learning Automata*. Englewood Cliffs, NJ: Prentice Hall.

Siskind, J.M. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition 61(1-2)*, 39-91.

Soderstrom, M., Seidl, A., Nelson, D., & Jusczyk, P. (2003). The prosodic bootstrapping of phrases: Evidence from prelinguistic infants. *Journal of Memory and Language, 49,* 249-267.

Yang, C. (2002). *Knowledge and learning in natural language.* Oxford: Oxford Univesity Press.

Yu, C. & Ballard, D. (2007). A unified model of early word learning: Integrating statistical and social cues. *Neurocomputing 70,* 2149-2165.