Using salience and hypothesis evaluation to learn object names in real time

Jon Stevens*

1 Introduction

The problem of how children learn the meanings of their first words, a problem for philosophers at least since the time of Augustine, has become an object of scrutiny in psychology and computational cognitive science. On one hand, experimental research shows us that young children can use a variety of cues (Bloom 2000) to learn meanings from context with only a few exposures (Carey 1978); on the other hand, computational modeling work underlines the difficulty of the process, requiring either complex statistical algorithms (Yu and Ballard 2007, Frank, Goodman, and Tenenbaum 2009) or large amounts of data (Fazly, Alishahi, and Stevenson 2010) to achieve adequate learning. The goal of this paper is to simplify the computational problem of early word learning by integrating empirically motivated cues into a simple statistical model that learns object names in real time with speed and precision.

Following Yu and Ballard 2007, we integrate prosodic and gestural cues into a statistical learning algorithm for object names (which comprise the bulk of an early child's vocabulary in many languages, including English). Unlike other models, we give the learner access to a lexicon that adaptively changes as new observations are processed. This allows the learner to check hypothesized word meanings against new input and to enforce a preference for one-to-one mappings between words and objects. These enrichments have roots in experimental research and allow us to construct a simple, effective, and principled online learning model based on rewarding and penalizing probabilities (following Yang 2002) associated with semantic hypotheses. The result is a psychologically plausible model that outperforms similar models (Fazly et al. 2010) and is competitive against more complex, implausible models (Frank et al. 2009).

2 Previous Work

Children are able to learn words from context, often taking entire sentences as input and breaking that input down to create word-to-meaning mappings. In addition to this task, which is far from trivial, children must filter out an infinite number of erroneous but logically possible hypotheses of word meaning, as Quine (1960) famously noticed. The quantum leap between considering each member of an infinite set (an impossible task) and considering each member of a finite set, no matter how large, is the basis for the claim that word learning relies fundamentally on innately given hypothesis space constraints. The question, then, is not whether learning is constrained, but how it is constrained. We take computational models to be tests of purported answers to this question. As such, models should reflect both the representations and the mechanisms present in human learners.

2.1 Experimental Work

We make use of three principles of early word learning that have emerged from experimental research: (1) mutual exclusivity, (2) the availability of gestural and prosodic cues, and (3) the apparent ability of learners to evaluate hypothesized word meanings against new data.

Markman (1992) and others have proposed that word learning is guided by a mutual exclusivity assumption, a default assumption that objects have only one name. There is independent experimental evidence (Ichinco, Frank, and Saxe 2009) that suggests that children disprefer many-to-one word-to-object mappings, and such a preference improves the performance of a simple learning model, excluding would-be distractors from the semantic hypothesis space when those distractors already have a name in the learner's lexicon. Markman's view is that mutual exclusivity acts in

*Thanks to Charles Yang, John Trueswell, Tamara Medina, and the attendees of PLC 35 for their input. All mistakes are my own.

concert with other default assumptions to extract a finite hypothesis space from Quine's infamous infinity.

Not only must the learner's hypothesis space be made finite, but it must interact with the learning mechanism in a way that produces quick results. Since Carey 1978 it has been noted that children learn words with impressive speed, often after only a few exposures. To achieve this end, we hold that word learning is guided not only by constraints like mutual exclusivity, but also by principles of salience and knowledge. This view allows the learning algorithm itself to be quite simple.

Under our conception of the process, word learning is guided both by word stress and by gestures, with greater weight being given to semantic hypotheses that map stressed words to gesturally indicated objects. Together we call these two cues "salience cues", reflecting their function of highlighting particularly important words and objects and making them salient to the learner. Without these crucial components, the data is simply too noisy for a simple learner to navigate. But these cues are independently justified. It is well known that babies are attentive to eye gaze and gestures. By nine months, they are capable of joint attention (Baldwin 1991, Bloom 2000), even responding to the emotional reactions of others. In short, humans seem to be programmed to pay attention to the actions of other humans from an early age. Thus, a gesture can serve as an "attentional magnet" for a young word learner.

If gesture serves to draw attention within the visual field, then patterns of prosodic prominence can be thought of as auditory gesture. Since the prosodic peaks of natural language have audible acoustic correlates (which are exaggerated in infant-directed speech), and since babies are known to be sensitive to these correlates (Soderstrom, Seidl, Nelson, and Jusczyk 2003, Thiessen, Hill, and Saffran 2005), we can posit that phonological phenomena such as word stress can be brought to bear on the question of how young learners figure out which words in an utterance are meant to refer. Indeed, prosodic information has been shown to be a good guide to word segmentation (Yang 2004), an ability that must precede word learning.

Finally, recent work suggests that word learning involves a form of hypothesis evaluation, whereby learners will guess at a word's meaning and then, as further utterances of that word are processed, search the object space for evidence supporting their guess. Medina, Trueswell, Snedeker, and Gleitman (2009) assess mechanisms of cross-situational learning in adults using the human simulation paradigm (Gillette, Gleitman, Gleitman, and Lederer 1999), a method whereby subjects are given video vignettes of word learning instances with the audio track removed and a single nonsense word uttered in place of some real word. Subjects were asked to give their best guesses as to the meaning of the nonsense words uttered in the vignettes. The vignettes were divided into "high informative" (HI) and "low informative" (LI) vignettes. The HI vignettes were those which were guessed correctly a majority of the time in isolation (determined in a separate experiment), and everything else was coded as a LI vignette.

Interestingly, subjects who saw a HI vignette followed by four LI vignettes were more likely to guess word meanings correctly at the end of the experiment than subjects who saw the same five vignettes in a different order. The authors hypothesize that early low informative instances handicap the learner, because rather than using the high informativity of later instances to make correct guesses, learners instead waste their time checking and rejecting the erroneous guesses they made previously. Subsequent studies show similar effects (Medina, Hafri, Trueswell, and Gleitman 2010). Subjects behave as if they are choosing a hypothesized meaning for a novel item, and then verifying or falsifying that meaning as new data is received. This process of hypothesis evaluation opposes the traditional view of cross-situational word learning as a process of associating words with sets of multiple co-present objects. The computational model presented here reflects these developments; we show that it is helpful for the learner to be able to evaluate the semantic hypotheses contained in their lexicon against new data.

2.2 Previous Models

Beginning with Siskind 2000, computational modeling has been a valuable tool for investigating the word learning process. Various approaches have been taken, including Bayesian inference (Niyogi 2002, Xu and Tenenbaum 2007, Frank et al. 2009) and machine translation (Yu and Ballard 2007,

Fazly et al. 2010).

Yu and Ballard's (2007) work is particularly interesting for our purposes because it demonstrates the positive effects that prosodic and gestural cues can have on model performance. A machine translation algorithm (Brown, Della Pietra, Della Pietra, and Mercer 1993) serves as a purely statistical core which is expanded by external social factors. The authors code corpus data for both prosodic peaks and indication by gesture or eye gaze. The words that represent peaks on an utterance's pitch track are given more weight than the other words in the utterance, and objects that are judged to be indicated in the visual field are given an analogous boost. We use a similar coding method. Yu and Ballard's is a batch learning model, which is less plausible than an online model in that it processes an entire corpus at once rather than learning words in real time. Fazly et al. (2010) remedy this by implementing the translation algorithm incrementally. It is this incremental variant that we use as a comparison model.

One of the most powerful recent models is the Bayesian model of Frank et al. 2009. Using Bayesian inference, this model assigns a posterior probability score to individual lexicons given a corpus of data. MCMC stochastic search is used to find the lexicon with the highest score; no claims are made about how human learners do this. The scoring algorithm considers all possible intended sets of referents for a given scene. For example, if two objects, a pig and a horse, are visible to the learner during a particular utterance, four possible intentions must be considered: the speaker could be talking about the horse, the pig, both, or neither. Each possible intention yields some probability value, and those values are added together to obtain the contribution of that utterance to a lexicon's overall score.

Although the lack of explicitly given clues about speaker intent is perceived as an advantage, there is no indication that this reflects the behavior of human learners. Furthermore, considering all possible intents adds considerable complexity to the model in that the lexicon scoring algorithm becomes exponentially more demanding the more cluttered the room is. Since values are computed over the power set of visible objects, a naming event involving n candidate objects will contribute 2^n calculations to the scoring process. This is not too problematic with relatively clean data, but one can easily imagine a naturalistic learning environment with 30 distinct objects in the visual field, which would require over a billion calculations just to score one lexicon.

Frank et al. claim that Bayesian inference explains mutual exclusivity. However, it is a choice by the modelers to make the likelihood term of their probability calculation dependent on the conditional probability $P(word|object)$, rather than $P(object|word)$. Thus, mutual exclusivity is built into the inference mechanism, not explained by it. Similarly, the machine translation algorithm used in Yu and Ballard 2007 and Fazly et al. 2010 is designed specifically to favor one-to-one mappings. In the absence of a deep explanation, we treat ME as an external cue rather than an architectural fact.

3 Model Overview

Word learning is mediated by a probability matrix with word types on the vertical axis and object types on the horizontal axis, illustrated in Figure 1.

The semantic hypothesis space for a potential object name is both open-ended and contingent on observation. This means that:

- A word-to-object mapping gets a value if and only if the word and the object have co-occurred.
- New words and objects can be introduced into the matrix at any time.

For example, the words *can*, *read*, and *books* are never uttered in the presence of the object coded 'EYES' in our evaluation corpus. Therefore, mappings from these words to 'EYES' have no value in Figure 1. This has the effect of reducing the size of a word's hypothesis space and preventing completely unfounded mappings from receiving a positive value when other mappings are penalized.

Novel words are mapped to 'NULL' with probability 1, with co-occurring objects receiving a value of 0. The 'NULL' mapping corresponds to the hypothesis that a word does not refer to an object. We take this to be the learner's default assumption. New objects are introduced into an old word's hypothesis space with a probability value of $\frac{1}{n}$, where n is the new size of that word's

	BOOK	BIRD	RATTLE	FACE	EYES	NULL
look	0.45	0.00	0.01	0.38	NA	0.16
we	0.00	0.01	0.00	0.01	NA	0.98
can	0.00	0.01	0.00	0.01	NA	0.98
read	0.01	0.00	0.01	0.00	NA	0.98
books	0.23	0.00	0.00	0.36	NA	0.41
david	0.36	0.00	0.00	0.23	NA	0.41

Figure 1: A partial probability matrix for words and objects

hypothesis space. The rest of the probability vector is normalized to accommodate the addition. This gives new semantic hypotheses a fair shot at lexicon inclusion.

This matrix provides us with a way to add and track probabilities of word-to-object mappings. Learning proceeds by updating these probabilities. We use Bush and Mosteller's (1951) Linear Reward-Penalty (LR-P) scheme, which was first applied to linguistic learning by Yang (2002). Below are the LR-P functions for rewarding and penalizing the probability of a hypothesis.

Table 1: Linear Reward-Penalty functions for a hypothesis h.

REWARD(h)	$p(h) = p(h) + \gamma(1 - p(h))$ where γ is some constant between 0 and 1 For all $h\prime \neq h$: $p(h\prime) = p(h\prime) * (1 - \gamma)$
PENALIZE(h)	$p(h) = p(h) * (1 - \gamma)$ For all $h\prime \neq h$: $p(h\prime) = \frac{\gamma}{n-1} + p(h\prime) * (1 - \gamma)$ where n is the number of hypotheses being considered

The learning coefficient γ determines the severity of rewards and penalties. The final version of our model uses variable γ values to represent the privileged status of salient words and objects. Using these functions we update probabilities on the fly, and we use the results to update the learner's current lexicon of word-object pairs by including all and only those pairs whose probability values exceed a given threshold. This threshold (set to 0.65 in our simulations) serves to transform the probabilities into a discrete set of mappings that the learner can evaluate.

We add the hypothesis evaluation component by treating words that are in the current lexicon differently than other words. If a word is already mapped to an object, then the probability associated with that mapping is rewarded or penalized depending on whether that object is in the present situation (i.e. depending on whether the learner's hypothesis is consistent with current observation). In this case, no other candidates are rewarded. In all models, mutual exclusivity is enforced by exempting objects that already have names from multiple-candidate rewarding. The algorithm is given in Figure 2.

The final component of our model is the integration of the salience cues. Objects in our video corpus were coded for gesture. An object was considered to be indicated by gesture during an utterance if it any point it was both (1) judged to be in the baby's field of vision, and (2) pointed to or held up in front of the baby. Eye gaze, being less obvious in the videos and therefore more prone to errors, was not coded.

Words were coded for prosodic accent. Utterances were given prosodic grid structures like the one in Figure 3, representing peaks in stress. Any word that received stress above the lexical level

For each observation, consisting of an utterance U and a randomly-ordered set of possible object referents O:

For each word w in U:

1. If w is novel, assign probability 1 to $w \rightarrow NULL$
2. Else, add new objects to w's hypothesis space.
3. If w is in the current lexicon:
   - $\Rightarrow$ If w's hypothesized meaning m is an element of O, reward$(w \rightarrow m)$.
   - $\Rightarrow$ Else, penalize $(w \rightarrow m)$.
4. If w is not in the current lexicon:
   - $\Rightarrow$ For each o in O:
     - $\Rightarrow$ If o is not in the current lexicon, reward$(w \rightarrow o)$.

Update the current lexicon.

Figure 2: An online cross-situational learning algorithm
[The arrow ($\rightarrow$) in the algorithm should be read "maps to".]

		x			
		x	x		x
x	x	x	x .	x	x .
There's	a	**bear**	**looking**	at	**David**

Figure 3: Stress on a prosodic grid

was coded as a stressed word. In typical adult speech the acoustic correlates of stress are subtle, and thus coding in this way is prone to subjectivity. However, this problem is ameliorated here, at least in part, by the exaggerated pronunciations utilized in the child-directed speech in the evaluation corpus.

Information about stress and gesture is used to determine the value of the learning coefficient γ for each rewarding or penalizing event. We give the model three parameters:

- γ_H is the learning coefficient used when rewarding or penalizing a mapping that is already in the lexicon (hypothesis evaluation).
- γ_M is the default learning coefficient used when rewarding possible mappings that are not already in the lexicon (multiple-candidate rewarding).
- b determines how much weight is given to hypotheses that map stressed words to gesturally indicated objects during multiple-candidate rewarding.

For words already in the lexicon, single hypotheses are rewarded or penalized with $\gamma = \gamma_H$. For words not in the lexicon, multiple possible mappings are rewarded with a different gamma value; mappings between stressed words and gesturally indicated objects are rewarded with $\gamma = \gamma_M * b$, while other mappings are rewarded with $\gamma = \gamma_M * (1 - b)$. The best performance is achieved when γ_H and γ_M are relatively high (0.4 and 0.36, respectively), and when most of the weight is given to salient mappings ($b = 0.98$).

To restate, the learner rewards and penalizes more drastically when checking their current lexicon against the world than when making multiple associations, and when the learner is making multiple associations, more weight is given to hypotheses that map stressed words to gesturally indicated objects.

uttered:	{there's, a, bear, looking, at, david}
stressed:	{bear, looking, david}
visible:	{BOOK, BIRD, RATTLE, BEAR, BOTTLE}
indicated:	{BEAR}
lexicon:	{david → MIRROR}

Figure 4: Example stimulus and accompanying lexicon

To illustrate, consider the utterance in Figure 3. Assume, as shown in Fig. 4, that there are five visible objects accompanying this utterance, and only one of them is indicated by gesture (the mother is pointing to the bear and ignoring the other objects). Upon hearing this utterance, the learner possesses a lexicon of one entry: the word "david" maps erroneously to the object 'MIRROR'.

These data will be processed incrementally by the learner in the following way:

1. Since *there's* is not in the lexicon, it undergoes multiple-candidate rewarding rather than single hypothesis evaluation. Since it is not stressed, all present object meanings are rewarded using the coefficient $\gamma_M * (1 - b)$.

2. The unstressed article *a* undergoes the same process as *there's*.

3. The lexicon does not have a mapping for *bear*, so it undergoes multiple-candidate rewarding, but since *bear* is stressed, the learning coefficient can vary. The gesturally indicated object referent 'BEAR' is rewarded with the higher coefficient $\gamma_M * b$, while the other non-indicated objects are rewarded with $\gamma_M * (1 - b)$.

4. The stressed verb *looking* undergoes the same process as *bear*.

5. The unstressed preposition *at* behaves like *there's* and *a*.

6. Since *david* has a mapping in the learner's current lexicon, only that mapping is considered. In this case, *david* maps to 'MIRROR', and the object 'MIRROR' is not present in the current scene, so the learner's hypothesis is penalized. If the penalty lowers the probability value below the given threshold, then *david* → 'MIRROR' is kicked out of the lexicon.

4 Performance and Comparisons

All models were run on hand codings of two videos of mother-child interaction from the Rollins corpus (CHILDES, MacWhinney 2000). Together the videos consist of 496 utterance-situation pairs (about 20 minutes of video). Performance was evaluated by aggregating the precision and recall against a gold standard over 100 simulations[1], and taking the harmonic mean of the average precision and recall to produce an F-score. Model performance is detailed in Table 2. Two online models were tested: one with a fixed γ value, and one which uses stress and gesture to determine γ. These models are compared to two implementations of Frank et al.'s (2009) Bayesian model: a direct implementation and a variant that only computes over stressed words and indicated objects.[2] We also compare performance to that of the less powerful but more realistic incremental machine translation model of Fazly et al. 2010, with and without external cues as implemented in Yu and Ballard 2007.

[1]Multiple simulations account for slight variations in output caused by randomizing the order in which multiple candidates are rewarded.

[2]We used our own hand-coding of the same videos that were used by Frank et al. For the Bayesian implementations, the authors' original code was used, strongly suggesting that the discrepancy between the performance reported here and the performance reported in Frank et al. 2009 is due to differences in the coding of the data.

Table 2: Model performance comparison.

Model type	Precision	Recall	F-score
Bayesian (FGT 09)	0.36	0.29	0.32
Bayesian (FGT 09) + stress and gesture	0.72	0.38	0.52
MT (FAS 10)	0.23	0.09	0.13
MT (FAS 10) + stress and gesture	0.29	0.55	0.35
Current	0.36	0.06	0.10
Current +stress and gesture	0.92	0.32	0.48

Word	Object	Word	Object
book	**book**	**piggies**	**pig**
bear	**bear**	**hat**	**hat**
bunny	**bunny**	**moocow**	**cow**
kittycat	**cat**	meow	cat
sheep	**sheep**	**bigbird**	**bird**
bird	**duck**	**ring**	**ring**

Figure 5: Most frequent output lexicon

We see that adding prosodic and gestural information is a boost to both types of models; however, the cues have the most drastic effect on the current model. Once the cues are integrated, the F-score is comparable to that of the Bayesian model, and considerably higher than that of the machine translation model. Though the Bayesian model achieves a slightly higher F-score, the current model has a decided advantage in precision, with almost no erroneous mappings remaining in the lexicon. The majority of simulations using this model produce the lexicon seen in Figure 5.

Performance is comparable to the Bayesian model of Frank et al. 2009, and our online learning model represents a computational simplification. Beal and Roberts (2009) argue for the importance of complexity analysis in computational cognitive science. A cognitive model should operate within known limits of human computational power, and complexity analysis is necessary to evaluate how realistic a model could be. Beal and Roberts show the Bayesian model of Xu and Tenenbaum 2007 to be quite costly from this perspective. Frank et al.'s model is even more costly. As mentioned above, it is problematic to sum probabilities for all possible intention sets for each situation. If the number of objects seen at one time has some upper bound N, then the upper bound asymptotic complexity will be $O(2^N)$; the time it takes to process one situation will grow exponentially with the number of visible objects. This is not a problem for relatively clean data like the videos from the Rollins corpus, where the number of visible objects does not typically exceed 6 or 7, but an especially cluttered room may force the learner to make billions of calculations to score one lexicon against one interaction. This problem does not arise in our model.

Finally, the model presented here holds the promise of further unification with experimental findings. Ongoing research is investigating the viability of a model that updates the probability of a single candidate meaning for each uttered word, never attending to multiple co-present meanings.

5 Conclusion

We have presented a model of object name learning that relies on gestural and prosodic cues and utilizes both single-candidate and multiple-candidate probability updating mechanisms. The model operates in real time, making only one pass through a corpus and updating a lexicon after each

successive utterance-situation pair. Performance is close to that of a comparable Bayesian model and better than that of a similar online model. The next step in this line of research is to link up this computational approach even closer with experimental findings, and it is our hope that in doing so we may contribute to the growing pool of knowledge about how children learn the meanings of their first words.

References

Baldwin, Dare. 1991. Infants' contribution to the achievement of joint reference. *Child Development* 62.

Beal, Jacob, and Jennifer Roberts. 2009. Enhancing methodological rigor for computational cognitive science: Computational complexity. *Cognitive Science Conference* .

Bloom, Paul. 2000. *How Children Learn the Meanings of Words*. Cambridge, MA: MIT Press.

Brown, Peter, Stephen Della Pietra, Vincent Della Pietra, and Robert Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics* 19:263–312.

Bush, Robert, and Frederick Mosteller. 1951. A mathematical model for simple learning. *Psychological Review* 68:313–323.

Carey, Susan. 1978. The child as word learner. In *Linguistic Theory and Psychological Reality*, ed. M. Halle, J. Bresnan, and G. Miller. Cambridge, MA: MIT Press.

Fazly, Afsaneh, Afra Alishahi, and Suzanne Stevenson. 2010. A probabilistic incremental model of word learning in the presence of referential uncertainty. *Cognitive Science* 34:1017–1063.

Frank, Michael, Noah Goodman, and Josh Tenenbaum. 2009. Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science* 20.

Gillette, Jane, Henry Gleitman, Lila Gleitman, and Anne Lederer. 1999. Human simulations of vocabulary learning. *Cognition* 73:135–176.

Ichinco, Denise, Michael Frank, and Rebecca Saxe. 2009. Cross-situational word learning respects mutual exclusivity. *Proceedings of the 31st Annual Meeting of the Cognitive Science Society* .

MacWhinney, Brian. 2000. *The CHILDES Project: Tools for Analyzing Talk*, volume 2. Mahwah, NJ: Erlbaum, 3 edition.

Markman, Ellen. 1992. Constraints on word learning: Speculations about their nature, origin, and domain specificity. In *Modularity and Constraints on Language and Cognition: The Minnesota Symposium on Child Psychology*, ed. M. Gunnar and M. Maratsos. Mahwah, NJ: Erlbaum.

Medina, Tamara, Alon Hafri, John Trueswell, and Lila Gleitman. 2010. Propose but verify: Fast-mapping meets cross-situational word learning. *Boston University Conference on Language Development* .

Medina, Tamara, John Trueswell, Jesse Snedeker, and Lila Gleitman. 2009. Rapid word learning under realistic learning conditions. *LSA Annual Meeting, San Francisco, CA* .

Niyogi, Sourabh. 2002. Bayesian learning at the syntax-semantics interface. *Proceedings of the 24th Annual Conference of the Cognitive Science Society* 697–702.

Quine, W.V.O. 1960. *Word and Object*. Cambridge, MA: MIT Press.

Siskind, Jeffrey. 2000. Learning word-to-meaning mappings. In *Models of Language Acquisition*, ed. P. Broeder and J. Murre. Oxford: Oxford University Press.

Soderstrom, Melanie, Amanda Seidl, Deborah Nelson, and Peter Jusczyk. 2003. The prosodic bootstrapping of phrases: Evidence from prelinguistic infants. *Journal of Memory and Language* 49:249–267.

Thiessen, Erik, Emily Hill, and Jenny Saffran. 2005. Infant directed speech facilitates word segmentation. *Infancy* 7:49–67.

Xu, Fei, and Josh Tenenbaum. 2007. Word learning as bayesian inference. *Psychological Review* 114.

Yang, Charles. 2002. *Knowledge and Learning in Natural Language*. Oxford: Oxford University Press.

Yang, Charles. 2004. Universal grammar, statistics, or both? *TRENDS in Cognitive Sciences* 8:451–456.

Yu, Chen, and Dana Ballard. 2007. A unified model of early word learning: Integrating statistical and social cues. *Neurocomputing* 70:2149–2165.

Department of Linguistics
University of Pennsylvania
Philadelphia, PA 19104-6305
jonsteve@ling.upenn.edu