How to Make the Most out of Very Little

Charles Yang*
University of Pennsylvania

March 2017

Pet ownership is at all time high. According to a 2006 Gallup survey, 44% of Americans owned a dog, 29% owned a cat, and 17% owned both. By 2012, the American Pet Products Association could reveal that these numbers had risen to 47% for dogs and 37% for cats. And multiple pet ownership had seen the most vigorous growth: a whopping 29% of Americans now own a dog as well as a cat.

For students of word learning, especially those under Lila Gleitman's tutelage, this last statistic should set off alarm bells. In a cluttered family room, where all members of the household congregate, how does the baby know that "dog" means DOG and "cat" means CAT, when both are likely to be present when either word is heard?[1] This is not even to consider the interference from CHAIR, CRAYON, DESK, RUG, SHOE, SPOON, WINDOW, etc., that are also in the mix (Medina et al. 2011).

In this short note, I provide a preliminary analysis of how the child may overcome the ambiguity problem to find the meanings of words. Drawing from Lila and her colleagues' research, I first review the severity of referential ambiguity in realistic word-learning environments (Section 1). Section 2 presents some current approaches to the problem, including the Pursuit model which arose from a recent collaboration (Stevens et al. 2016). Surprisingly, resource-limited models that entertain a subset of available word-meaning pairings outperform "big data" models that attend the totality of learning experience. Section 3 sketches out a formal account for this paradoxical result. I suggest that the statistical distribution of language use, especially the sparsity of highly informative instances, in fact favors the restrictive models of word learning.

1 Of Pictures and Words

As Lila and Henry Gleitman once memorably remarked (1992), a picture is worth a thousand words, and *that's the problem*. Language is much more than here-and-now, and words do not present themselves in the environment to be readily picked up. When we say "let's write a letter to grandma", the "letter" is still imaginary and grandma may well be tending her garden in Canada. Indeed, the noisy pairing between words and their meanings, and more generally the creative aspect of language use, form a core argument in Chomsky's critique of Skinner's associationist

*Thanks to my co-conspirators Lila Gleitman, John Trueswell, and Jon Stevens. All errors are my own.

[1] I follow the Fodorian notion of using uppercase to denote the concept/referent (e.g., DOG) picked up by a phonological word (e.g., "dog").

program (1959). Yet children learn words rapidly and accurately (Carey and Bartlett 1978, Miller 1991, Bloom 2000), and in strikingly uniform ways (Landau and Gleitman 1985).

No one can seriously question the role of the experience in language acquisition: after all, New York kids learn English ("dog") and Beijing kids learn Chinese ("gou"). The Human Simulation Paradigm (Gillette et al. 1999) has proved an effective tool to assess the ecological condition of language acquisition, and the environmental challenges children have to overcome to learn the meanings of words. In a typical study, adult participants watch silent videos of young children interacting with their mothers during. They hear a beep when a target word is produced by the caretaker. Even though the participants are competent speakers who already have a full vocabulary, their performance is abysmal. On average, they could only identify nouns with 45% accuracy; for verbs, the accuracy drops to 15%. While modest improvement for nouns can be observed across multiple learning instances there is hardly any benefit for the verbs. The degree of referential ambiguity is very high: it is very difficult for a learner to correctly identify words solely from observation even though it's quite not as bad as choosing one out of a thousand.

But there is hope. One of the consistent findings from the Human Simulation Paradigm is that while most words are hard to pick out from the environment, some are relatively easy. Conventionalized expressions ("Hi", "Bye-bye") and concrete nouns ("ball", "plane", "swing"), including those representing basic-level categories (Gentner 1982), are relatively easy to identify and may well serve as the foundation for vocabulary as well as grammatical acquisition (Pinker 1984, Gleitman et al. 2005). In the most recent and most comprehensive study (Trueswell et al. 2016), the Human Simulation Paradigm was taken out of the laboratory to assess the potency of observational learning in the naturalistic situation of language acquisition. Using a video corpus of 351 vignettes of free play and focusing only on concrete nouns, Trueswell et al. 2016 find that, on average, only 18% of the guesses are correct but a good deal of variance is found within. In particular, there is a small but non-negligible subset of vignettes, about 15% in all, where participants guess the mystery noun correctly more than 70% of the time. The high informativity of these scenes appear to be the result of joint attention (Tomasello and Farrar 1986, Baldwin 1991, Woodward 2003) and clear visual signatures (Pereira et al. 2014) especially temporal cues indicative of causation (Scholl and Tremoulet 2000).

So Mother Nature may be parsimonius but she is not mean. For children, the smart move is to make maximum use of these highly informative learning instances. Such instances are identified by relying on heuristics and constraints to narrow down the range of word-meaning mappings. As just reviewed, these cues may be social, attentional, perceptual in nature and potentially domain general. They may also be structural and reflect domain-specific properties of human language (Markman 1990, Markman et al. 2003, de Marchena et al. 2011), especially in the study of syntactic boostrapping (Gleitman 1990, Naigles 1990, Fisher et al. 1994, Gillette et al. 1999).

This paper tries to characterize word learning mechanisms that can make the most out of the rare but highly effective learning instances. For ease of exposition, I will focus on how to learn that "dog" means DOG. The problem is complicated by CAT, which is almost always present when the word "dog" is uttered. But every now and then the DOG goes out for a walk and the child comes along: Crucially, the CAT stays home and is therefore not in contention. Thus, the alone time with the DOG becomes a highly informative learning instance: say, 15% of the time when "dog" is uttered, following the statistical findings by Trueswell et al. 2016. While this somewhat silly example focuses on the learning of concrete nouns, it represents a wide range of word learning situations. Taking the DOG for a walk is equivalent to the occasional pointing or eye gaze direction

that can guide children's attention, thereby disambiguating CUP from BALL on the kitchen table. Or it may represent a piece of morphosyntactic information ("a seb" vs. "sebbing") that settles the category of an unknown word (Brown 1957). DOG and CAT may also be stand-ins for CHASE and FLEE, PUSH and MOVE, etc., which are among Lila's favorite verbs. While these twin verbs are often both compatible with the observation of an event — "I'm *pushing* the box" and "I'm *moving* the box" — walking the DOG is formally equivalent to the child hearing a critical disambiguating syntactic frame (Gleitman 1990, Naigles 1996): we can say "the box is moving" but not "the box is pushing", "I'm pushing on the box" but not "I'm moving on the box", which point to the vital syntactic and semantic differences between these verbs (Levin 1993).

With the problem of referential ambiguity in mind, let's turn to some word learning models on the market.

2 Two Word Views and Three Learning Models

A picture may be worth a thousand words but what if each word is paired with a thousand pictures themselves? For example, the word "dog" may, on average, be accompanied with more learning instances that contain DOG than CAT, and will emerge as the winning candidate over time. The modern idea of cross-situational learning (Pinker 1994, Yu and Smith 2007) is an update of the associationist program for language (e.g., Osgood et al. 1957, Quine 1960). As long as the target meaning is associated sufficiently strongly — more strongly than its competitors — the learner may be able to tabulate and detect such statistical regularity. In the new age of empiricism fueled by Big Data and fast machines (Roy et al. 2006, Le 2013), this all seems plausible.

Unless children literally learn everything about a word in one shot, learning must make use of multiple learning instances. The key question concerns how such cross-situational information is used. It is instructive to distinguish two general approaches in the recent literature. A *global* approach (Siskind 1996, Yu and Smith 2007, Frank et al. 2009, Fazly et al. 2010, McMurray et al. 2012) resolves referential ambiguity by aggregating situational data for large number of word-referent co-occurrences. The meaning of a word is the candidate with the strongest statistical correlation over all learning instances. By contrast, a *local* approach (Medina et al. 2011, Trueswell et al. 2013, Stevens et al. 2016) attempts to resolve ambiguity in the moment. Specifically, it ignores all potential words meanings that do not serve to confirm or disconfirm a word's hypothesized interpretation. In the limit, a local model resolves ambiguity by ignoring ambiguity: It only learn from unambiguous data, a powerful idea that has desirable formal properties (Angluin 1980) and has been proposed for other problems in language acquisition (Fodor 1998).

To understand the differences between the local and global approaches to word learning, consider an illustrative example in Figure 1. The word is "mipen", meaning ELEPHANT, and it is presented in five learning instances over time. Note that in learning instance (d), the target ELEPHANT is missing. This contrasts with almost all cross-situational learning experiments where the target meaning is always present, but is designed to better reflect language use in the real world, which is not bound by the here-and-now.

Consider the simplest global learner model, one which tabulates all word-object pairings and selects the winner in the end; the computational model of Fazly et al. 2010 is a faithful implementation of this idea. Figure 2(a) represents the internal state of the model during each learning instance, where each meaning candidate increases its score by one whenever "mipen" is heard.

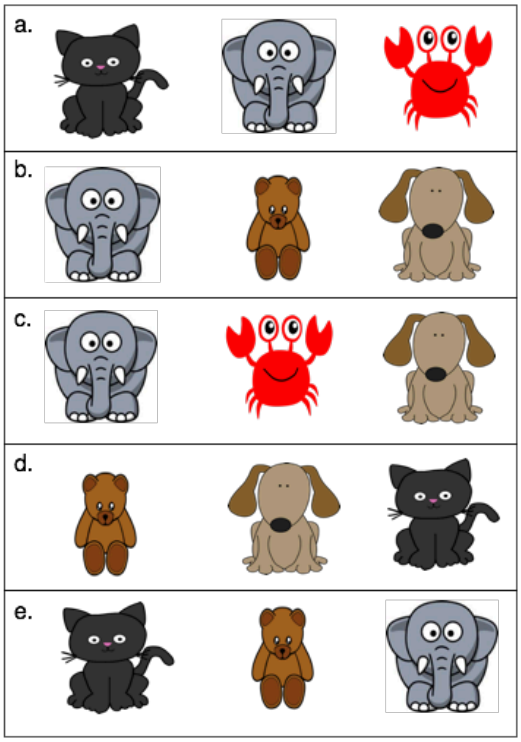
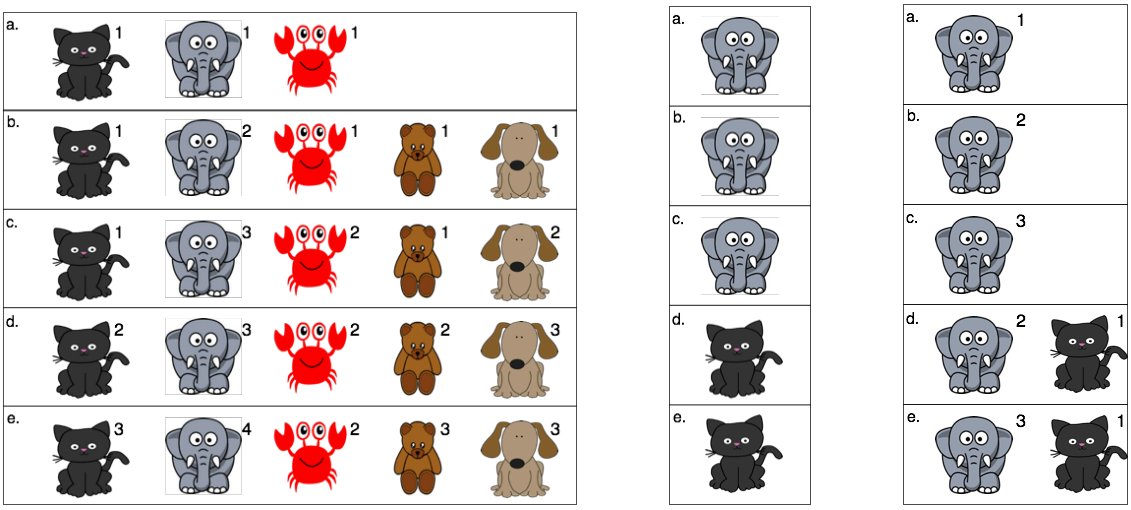


Figure 1: Mipen. Adapted from Stevens et al. 2016.



(a) Cross-situational learning (b) PbV (c) Pursuit

Figure 2: The dynamics of three word learning models. Adapted from Stevens et al. 2016.

Despite the absence of ELEPHANT in learning instance (d), it still emerges the winning candidate at the conclusion of learning.

The Propose-but-Verify model (PbV; Medina et al. 2011, Trueswell et al. 2013) is the simplest local model and it operates in the tradition of hypothesis testing. The learner maintains a hypothesis for a word and check it against each instance of the input data. The hypothesis is retained if confirmed (e.g., the referent is observed); otherwise it is replaced by a new hypothesis (e.g., another referent from the environment). The dynamics of PbV is shown in Figure 2(b). For the sake of argument, suppose the learner initially guessed that "mipen" is correct. This hypothesis is subsequently confirmed for learning instance (b) and (c) and is thus retained. However, it is jettisoned due to absence of ELEPHANT in (d), and another hypothesis from the observation — say, CAT — is selected. The updated hypothesis is subsequently confirmed and the learner concludes, incorrectly, that "mipen" means CAT. It is important to note that a PbV learner never considers more than two hypotheses — the old hypothesis, and a new one if the old is rejected — and all other possible referents are completely ignored; hence the very "narrow" internal states in Figure 2(b). Note also that the PbV model is stochastic and Figure 2(b) is just one of the many paths available to the learner. They could have guessed CRAB initially but stumble on ELEPHANT in the final instance.

The Pursuit model can be viewed as a probabilistic update of PbV. Here the hypotheses are stored in the memory with weights. Like PbV, this only applies to the hypotheses actively considered by the learner but not those that merely cooccur with the word: hence the slightly "wider" internal state in Figure 2(c) than PbV but much smaller than the cross-situational learner. Although there may be multiple hypotheses associated with a word, only one is tested against the input (again like PbV): its success or failure results in increment or decrement of the weights, following a simple scheme of reinforcement learning (Bush and Mosteller 1951, Rescorla and Wagner 1972, Sutton and Barto 1998, Dayan and Daw 2008) which has been effectively applied in other domains of language acquisition (Yang 2002, 2004). More specifically, Pursuit adopts a very aggressive "greedy" form of reinforcement learning. When multiple hypotheses are available in the memory, the learner will singlemindedly *pursue* the best hypothesis, one which has the highest score.[2] If confirmed, the rich will get richer. If it fails to be confirmed, its probability is decreased: the learner will add a new candidate to the memory (again like PbV) but the just defeated candidate may still remain the best.

Consider how Pursuit learns "mipen". Pursuit is also stochastic so its learning trajectory in Figure 2(c) is only one of the possible paths available to the learner. The score keeping of Pursuit is somewhat complex (see Stevens et al. 2016 for details) and we will use a simple counter here to illustrate the dynamics of learning. The initial (fortuitous) guess of ELEPHANT is rewarded continuously and reaches a score of 3. Note that since ELEPHANT has been confirmed, the learner does not register no other referents at all. At instance (d), however, the absence of ELEPHANT lowers its score 2. The learner add another hypothesis, say CAT, to the list with a score of 1. In

[2]Most instantiations of reinforcement learning sample among the hypotheses according to their probabilities. It so implemented, the probabilistic model becomes formally equivalent to an online version of cross-situational learning and will inherit all the empirical problems (Stevens et al. 2016). The question is why a greedy scheme is used in word learning while human subjects are clearly capable of sampling over multiple hypotheses in probability matching (Herrnstein and Loveland 1975) and in the evaluation of grammatical hypotheses (Yang 2002). I speculate that the use of Pursuit in word learning may reflect the structure of the lexicon. Certain frequency effects (e.g., Swinney 1979) are consistent with a list-like representation of lexical entries ranked by frequency (Murray and Forster 2004, Yang 2016).

the final instance, the learner *pursues* the best hypothesis, which is still ELEPHANT despite the earlier mishap. It restores its score to 3 and will be selected as the winner in the end.

At least in this example, Pursuit captures the best of both worlds: the computational simplicity of PbV and the effect of cumulative evidence from cross-situational learning. Occasional failures such as learning instance (d) in the "mipen" example will not be catastrophic. Furthermore, if "mipen" indeed means CAT, then the earlier misjudgment (ELEPHANT) presumably will be penalized in the future learning and eventually lose its lead. The memory component of Pursuit also accounts for the finding that hypotheses, if (and only if) entertained at some point during learning, leave behind a trace so that their probabilities of being selected in the end are greater than chance (Köhne et al. 2013).

Stevens et al. (2016) provide extensive computational analyses of the three models. The results include a surprisingly finding is that both local models provide at least as good, if not better, account of the behavioral results from Yu and Smith (2007), the paradigm study in support of global cross-situational learning. Experimental findings from other studies (e.g., Medina et al. 2011, Köhne et al. 2013, Trueswell et al. 2013) provide more direct and decisive support for the local models over global models.

Here I focus on another surprisingly result from our study, a quantitative evaluation of how these models fare "in the wild". To maintain continuity with previous modeling efforts (Yu and Ballard 2007, Frank et al. 2009), we annotated two video clips from the CHILDES database (MacWhinney 2000). A learning instance is defined by a child-directed utterance; it consists of the set of words used and the set of perceptually salient referents, judged by the annotator, at the time of the utterance. In addition to the Fazly et al. (2010) cross-situational model, the another global model (Frank et al. 2009) is also tested. After the entire corpus is processed, the models' output is compared against a gold standard lexicon of 34 words that are deemed plausibly learnable by the annotator; precision, recall, and F-score are then reported. As is conventional in computational modeling work, the free parameters for all models are manually tuned so as to produce the best performance possible. Again, the cross-situational model is deterministic and was only run once, while both PbV and Pursuit are stochastic and their performance figures are averaged over 1,000 simulations. Frank et al. (2009)'s Bayesian model attempts to produce the "best" lexicon given the distribution of words and referents in the learning data: since the number of possible lexicons is astronomically large, approximation methods were used to evaluate the model in a reasonable amount of time. The results are summarized in Table 1.[3]

These results can be only regarded as preliminary. In fact, they are frankly terrible: parents would be running to the family pediatrician if children were to learn more words incorrectly than correctly. Nevertheless, it is very surprising that Pursuit, a local model that considers a subset of word-referent pairings, is at least competitive with, if not better than, global models that optimize over the totality of the learning data. I now turn to investigate this puzzle.

[3]Our results are lower than those reported in Frank et al. (2009) on the same video corpus. There are two possible reasons for this. First, our annotation encodes a higher degree of ambiguity than that by those authors on the same corpus, which is still more ambiguous than that in the original Yu and Ballard (2007) study. On average, each referent in our annotation is paired with three words but with only two words in Frank et al. (2009)'s annotation. Second, the Bayesian model is computationally very expensive: to achieve the results reported in Frank et al. (2009) required five hundred hours of computer time (Michael Frank, personal communication), which was not practical given our computing facilities.

Model	Precision	Recall	F-score
Propose-but-Verify	0.04	**0.45**	0.07
Cross-situational	0.39	0.21	0.27
Bayesian	**0.50**	0.29	0.37
Pursuit	0.45	0.37	**0.41**

Table 1: Model performance on a small video corpus of child-directed English.

3 The Race between DOG and CAT

Let's revisit DOG vs. CAT, a toy example that is nevertheless emblematic of ambiguity resolution in word learning. As reviewed in Section 1, word-referent relations are is in general highly ambiguous but there are rare instances of clarity. For the present analysis, is not important what make these instances informative (see Section 1): it is sufficient to assume that they exist but with low probabilities. This allows us to use the DOG and CAT example to study the dynamics of word learning: DOG and CAT are generally indistinguishable but the occasional walk DOG takes in the absence of the CAT constitutes the critical disambiguating evidence. In what follows, I will make some suggestions about how to make the most out such precious information, and why local models such as Pursuit are best positioned to do so. If correct, it also goes some way toward making sense of the simulation results that models with severe resource limitation outperform models with the totality of data at their disposal.

To begin, I will make some simplifying assumptions about the learning situation and the learning models. These will make a formal analysis tractable while still shedding light on the core task of referential ambiguity resolution. Specifically, let's assume that there are N learning instances in which both DOG and CAT are present; for each instance, DOG and CAT have an equal probability of p_N of being selected. Thus, p_N is a measure of referential ambiguity; for instance, p_N may be 1.0 over the number of potential referents in the environment. These N instances are then followed by U instances where CAT is absent and DOG has a probability p_U of being selected each time. The Human Simulation Paradigm analyses reviewed in Section 1 suggest that in general, $U \ll N$ and $p_U \gg P_N$.

It is clear that DOG and CAT will not be differentiated in the first N learning instances. For the cross-situational learner, DOG and CAT will have the same numerical score. For Pursuit, which is stochastic, DOG and CAT are expected to be in a tie. We are thus interested in quantifying the effectiveness of the U instances as tie-breakers: the further DOG is "ahead" of CAT, the more likely is DOG learned as the meaning of "dog".

For a prototypical cross-situational learning (e.g., Fazly et al. 2010), the advantage of DOG over CAT is simply the proportion of the learning instances that contain DOG but not CAT. Let this term of advantage be Δ_X (X for cross-situational):

$$\Delta_X = \frac{U}{N + U} \tag{1}$$

I stress again that we are only considering the competition between DOG and CAT for the word "dog", thereby completely ignoring other potential referents that the learning model may have picked up. In the current setup, if the learner were to choose between DOG and CAT after the

N instances, they would be at 50% under all learning models. The term advantage is a quantitative measure of the expected "above chance" probability conferred by the subsequent U disambiguating instances.

Note that strictly speaking, a "true" cross-situational learner will always select DOG with probability 1 as long as $U > 0$ because DOG will have a higher score than CAT. But the calculation of the cumulative advantage afforded by U in Eq. 1 is important for several reasons. On the one hand, experiments (Smith and Yu 2008, Medina et al. 2011, Trueswell et al. 2013) clearly show that in cross-situational learning, the target meanings are only preferred above chance rather than categorically selected; see Stevens et al. (2016) for discussion. This suggests that a cross-situational learning model must have a probabilistic/scalar component that encodes the strength of statistical correlation such as Eq. 1. On the other hand, in the computational evaluation of word learning models, it is necessary to introduce a threshold function, similar to those used in other learning and memory models (e.g., Anderson and Schooler 1991), that determines whether a meaning has been successfully acquired (e.g., Frank et al. 2009, Fazly et al. 2010, Stevens et al. 2016). These threshold functions operate on the *relative* strength of the candidates. If the advantage of DOT over CAT is weak, then the threshold must be set low in order for DOG to be acquired, but doing so will let in more false positives in the overall assessment of the learning model. Thus, it is only safe to set the association threshold to a relatively high value to attain high accuracy necessary to model human word learning. This notion is again captured by Δ_X: the higher it is, the more likely will the cross-situational learner succeed to acquire DOG.

To be fair to the cross-situational approach, consider a more judicious learner. It still keeps track of all word-referent pairings but gives more credence to more informative learning instances. This reasonable assumption enables the learner to capitalize on low ambiguity instances: an observation with only one distractor is more useful than one with ten distractors. More concretely, let's introduce a "bias" factor for the U informative instances, and the resulting advantage for DOG over CAT for the *biased* cross-situational learner Δ_B is:

$$\Delta_B = \frac{UB}{UB + N}, \text{ where } B = \frac{p_U}{p_N} \tag{2}$$

Consider now Pursuit. For the formal analysis to be tractable, I will use a modified model of Pursuit that only counts confirmations but does not penalize for disconfirmations This simplification is justified because we are only comparing DOG and CAT and ignore other candidates the learner might have conjectured along the way. Because Pursuit always chooses the leader when learning stops, the problem becomes the calculation of the probability with which DOG is ahead of CAT after all $(N + U)$ learning instances. Since the first N instances offer no opportunity for DOG to beat CAT on average, the decisive advantage can only be provided by the U instances toward the end.

The race between DOG and CAT under Pursuit can be analyzed as a multinomial process which select each with probability p_N during the first N instances, following a binomial process that selects DOG with probability p_U in the subsequent U instances. There are three possibilities after N:

- DOG is already ahead of CAT: U will not provide any additional advantage.
- CAT is at least U steps ahead of DOG: DOG cannot be salvaged by U.

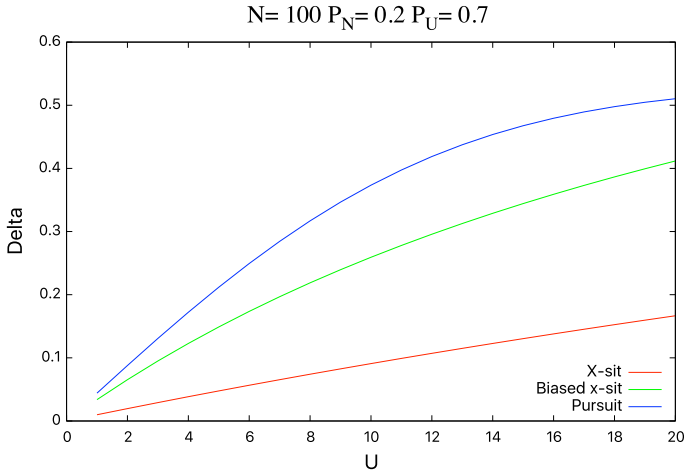


Figure 3: Pursuit makes better use of informative learning instances.

- CAT is i ($0 \leq i < U$) steps ahead: DOG can catch up if it is selected at least ($i + 1$) times during U.

Since Pursuit is a stochastic process, the first two events are just luck that has nothing to do with U. The probability of the third event is the unique contribution of U to push DOG ahead of CAT. Thus, the advantage for a Pursuit model (Δ_P) afforded by U is:

$$\Delta_P = \sum_{\substack{d \\ 0 \leq i < U}} \left[\binom{N}{d, d+i} p_N^d p_N^{d+i} (1 - 2p_N)^{N-2d-i} \cdot \sum_{j=i+1}^{U} \binom{U}{j} P_U^j (1 - P_U)^{U-j} \right] \tag{3}$$

The monstrosity in Eq. 3 is interpreted as follows. In the N instances, DOG has been chosen d times while CAT has been $d + i$ ($0 \leq i < U$) times and is thus i steps ahead. The first term inside the square brackets expresses this multinomial probability. The second term is the probability of DOG being selected j times in the U learning instances, where $j > i$ guarantees that DOG will surpass CAT in the end.

Let us turn to some numerical results. Following the Human Simulation Paradigm results (Trueswell et al. 2016), I will assume $P_N = 0.2$, which is roughly the average probability with which the subjects were able to identify the target referent, and $P_U = 0.7$, the probability of success on highly informative scenes. Figure 3 shows the Δ's as a function of U (from 1 to 20) for the three types of learner just considered, where N has been set to 100.

Figure 4 fixes the value of U at 15, again roughly following Trueswell et al. 2016, that approximately 15% of learning instances are highly informative. It explores the effect of the bias factor B (P_U/P_N), which measures the informativeness of U (over N). Again, we set $P_N = 0.2$ but values of B will range from 0.25 to 4, corresponding to $P_U = 0.05$ where U is actually less informative than N, and $P_U = 0.8$, where U is much more informative than N.

The formal analyses presented here can no doubt be enriched and refined but they already provide useful insight on the nature of word learning mechanisms. Both Figure 3 and Figure 4 suggest that a Pursuit-like model is much better equipped than cross-situational models to capitalize on

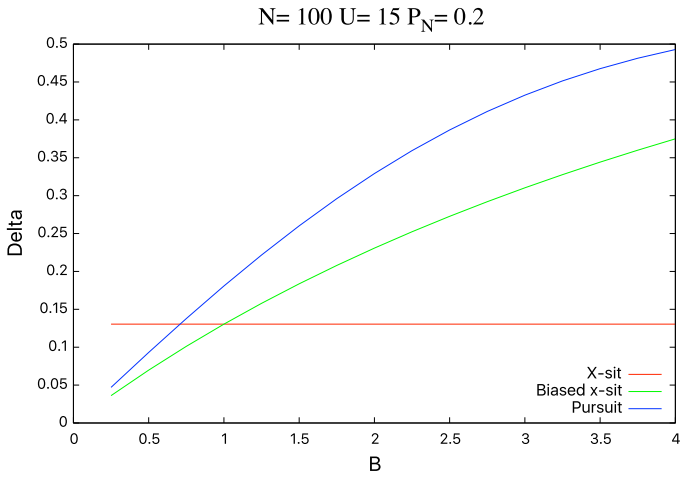


Figure 4: Pursuit makes better use of more informative learning instances.

the rare but highly informative learning instances. Since the model can only select one hypothesis, the probability of selecting the target meaning, already higher in the informative instances, increases sharply. This can effectively build a decisive advantage over its competitors. By contrast, a cross-situational model, even one which dynamically takes informativeness into account (Eq. 2), suffers from the problem of "dilution": informative learning instances are diminished in the regression to the mean, which is dominated by the vast number of largely useless data.[4] The signal is out there but it is faint: the worse thing that could happen is for it to be drowned out by the noise.

Children's resource limitation, then, may well turn out to be a blessing, which echos suggestions elsewhere in the developmental and learning literature (e.g., Newport 1990, Yang 2016). As Lila has taught us, word learning is hard; the best solution is not to try harder.

References

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

Angluin, D. (1980). Inductive inference of formal languages from positive data. *Information and Control*, 45(2):117–135.

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

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

[4] I do not wish to imply that a cross-situational learning model cannot make use of informative learning instances as effectively as Pursuit. For instance, the bias function in Δ_B (Eq. 2) can be modified to increase the benefit of the U instances exponentially. Nevertheless, I do maintain that it is attractive for the benefit to fall out of a simple learning model than for it to be implemented as an auxiliary component of a more complex one.

Brown, R. W. (1957). Linguistic determinism and the part of speech. *The Journal of Abnormal and Social Psychology*, 55(1):1–5.

Bush, R. R. and Mosteller, F. (1951). A mathematical model for simple learning. *Psychological Review*, 68(3):313–323.

Carey, S. and Bartlett, E. (1978). Acquire a single new word. *Child Language Development*, 15.

Chomsky, N. (1959). A review of B.F. Skinner's *Verbal Behavior*. *Language*, 35(1):26–58.

Dayan, P. and Daw, N. D. (2008). Decision theory, reinforcement learning, and the brain. *Cognitive, Affective, & Behavioral Neuroscience*, 8(4):429–453.

de Marchena, A., Eigsti, I.-M., Worek, A., Ono, K. E., and Snedeker, J. (2011). Mutual exclusivity in autism spectrum disorders: Testing the pragmatic hypothesis. *Cognition*, 119(1):96–113.

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

Fisher, C., Hall, D. G., Rakowitz, S., and Gleitman, L. (1994). When it is better to receive than to give: Syntactic and conceptual constraints on vocabulary growth. *Lingua*, 92:333–375.

Fodor, J. D. (1998). Unambiguous triggers. *Linguistic Inquiry*, 29(1):1–36.

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

Gentner, D. (1982). Why nouns are learned before verbs: Linguistic relativity versus natural partitioning. In Kuczaj, S. A., editor, *Language, thought, and culture*, pages 301–334. Lawrence Erlbaum Associates, Inc, Hillsdale, NJ.

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

Gleitman, L. (1990). The structural sources of verb meanings. *Language Acquisition*, 1(1):3–55.

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

Gleitman, L. R. and Gleitman, H. (1992). A picture is worth a thousand words, but that's the problem: The role of syntax in vocabulary acquisition. *Current Directions in Psychological Science*, 1(1):31–35.

Herrnstein, R. J. and Loveland, D. H. (1975). Maximizing and matching on concurrent ratio schedules. *Journal of the Experimental Analysis of Behavior*, 24:107–116.

Köhne, J., Trueswell, J. C., and Gleitman, L. R. (2013). Multiple proposal memory in observational word learning. In *Proceedings of the 35th Annual meeting of the Cognitive Science Society. Austin, TX: Cognitive Science Society.*

Landau, B. and Gleitman, L. R. (1985). *Language and experience: Evidence from the blind child*, volume 8. Harvard University Press, Cambridge, MA.

Le, Q. V. (2013). Building high-level features using large scale unsupervised learning. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 8595–8598. IEEE.

Levin, B. (1993). *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press, Chicago.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk*. Lawrence Erlbaum, Mahwah, NJ, 3rd edition.

Markman, E. M. (1990). Constraints children place on word meanings. *Cognitive Science*, 14(1):57–77.

Markman, E. M., Wasow, J. L., and Hansen, M. B. (2003). Use of the mutual exclusivity assumption by young word learners. *Cognitive psychology*, 47(3):241–275.

McMurray, B., Horst, J. S., and Samuelson, L. K. (2012). Word learning emerges from the interaction of online referent selection and slow associative learning. *Psychological review*, 119(4):831.

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

Miller, G. A. (1991). *The science of words*. Scientific American Library, San Francisco.

Murray, W. S. and Forster, K. I. (2004). Serial mechanisms in lexical access: The rank hypothesis. *Psychological Review*, 111(3):721–756.

Naigles, L. (1990). Children use syntax to learn verb meanings. *Journal of Child Language*, 17(02):357–374.

Naigles, L. R. (1996). The use of multiple frames in verb learning via syntactic bootstrapping. *Cognition*, 58(2):221–251.

Newport, E. (1990). Maturational constraints on language learning. *Cognitive Science*, 14(1):11–28.

Osgood, C. E., Suci, G. J., and Tannenbaum, P. H. (1957). *The measurement of meaning*. University of Illinois Press.

Pereira, A. F., Smith, L. B., and Yu, C. (2014). A bottom-up view of toddler word learning. *Psychonomic Bulletin & Review*, 21(1):178–185.

Pinker, S. (1984). *Language learnability and language development*. Harvard University Press, Cambridge, MA.

Pinker, S. (1994). How could a child use verb syntax to learn verb semantics? *Lingua*, 92:377–410.

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

Rescorla, R. A. and Wagner, A. R. (1972). A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In Black, A. H. and Prokasy, W. F., editors, *Classical conditioning II: Current research and theory*, volume 2, pages 64–99. Appleton Century Crofts, New York.

Roy, D., Patel, R., DeCamp, P., Kubat, R., Fleischman, M., Roy, B., Mavridis, N., Tellex, S., Salata, A., Guinness, J., et al. (2006). The human speechome project. In *Lecture Notes in Computer Science*, pages 192–196. Springer.

Scholl, B. J. and Tremoulet, P. D. (2000). Perceptual causality and animacy. *Trends in cognitive sciences*, 4(8):299–309.

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

Smith, L. and Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, 106(3):1558–1568.

Stevens, J., Trueswell, J., Yang, C., and Gleitman, L. (2016). The pursuit of word meanings. *Cognitive Science*, (To appear).

Sutton, R. S. and Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge University Press.

Swinney, D. A. (1979). Lexical access during sentence comprehension: (re)consideration of context effects. *Journal of Verbal Learning and Verbal Behavior*, 18(6):645–659.

Tomasello, M. and Farrar, M. J. (1986). Joint attention and early language. *Child development*, pages 1454–1463.

Trueswell, J. C., Lin, Y., Armstrong, B., Cartmill, E. A., Goldin-Meadow, S., and Gleitman, L. R. (2016). Perceiving referential intent: Dynamics of reference in natural parent–child interactions. *Cognition*, 148:117–135.

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

Woodward, A. L. (2003). Infants' developing understanding of the link between looker and object. *Developmental Science*, 6(3):297–311.

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

Yang, C. (2004). Universal grammar, statistics or both? *Trends in Cognitive Sciences*, 8(10):451–456.

Yang, C. (2016). *The price of linguistic productivity: How children learn to break rules of language*. MIT Press, Cambridge, MA.

Yu, C. and Ballard, D. H. (2007). A unified model of early word learning: Integrating statistical and social cues. *Neurocomputing*, 70(13–15):2149–2165.

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