TOPICS
TOPICS IN COGNITIVE SCIENCE

Topics in Cognitive Science (2019) 1–17
© 2019 Cognitive Science Society, Inc. All rights reserved.
ISSN:1756-8765 online
DOI: 10.1111/tops.12415

This article is part of the topic "2017 Rumelhart Prize Issue Honoring Lila Gleitman," Barbara Landau (Topic Editor). For a full listing of topic papers, see http://onlinelibrary.wiley.com/journal/10.1111/(ISSN)1756-8765/earlyview

How to Make the Most out of Very Little

Charles Yang

Department of Linguistics and Computer and Information Science, University of Pennsylvania

Received 15 March 2017; received in revised form 18 June 2018; accepted 9 July 2018

Abstract

I review the problem of referential ambiguity that arises when children learn the meanings of words, along with a number of models that have been proposed to solve it. I then provide a formal analysis of why a resource-limited model that retains very few meaning hypotheses may be more effective than "big data" models that keep track of all word-meaning associations.

Keywords: Word learning; Mathematical modeling; Cross-situational learning; Language acquisition

1. Introduction

Pet ownership keeps on rising. 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 grown to 47% for dogs and 37% for cats. 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, and so on that are also in the mix (Medina, Snedeker, Trueswell, & Gleitman, 2011).

Correspondence should be sent to Charles Yang, Department of Linguistics and Computer and Information Science, University of Pennsylvania, 3401 Walnut Street, 315-C Philadelphia, PA 19104. E-mail: charles.yang@ling.upenn.edu

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 the problem in realistic word learning environments (Section 2). Section 3 reviews some current approaches to the problem, including the Pursuit model that arose from a recent collaboration (Stevens, Gleitman, Trueswell, & Yang, 2017). Surprisingly, resource-limited models that entertain a subset of available word-meaning pairings outperform "big data" models that attend the totality of the learning experience. Section 4 sketches out a formal account of 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.

2. Gleitman's Problem

As Lila and Henry Gleitman once memorably remarked (1992), a picture is worth a thousand words, and *that is 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, forms a core argument in Chomsky's critique of Skinner's associationist program (Chomsky, 1959). Yet children learn words rapidly, accurately, and in strikingly uniform ways (Bloom, 2000; Carey & Bartlett, 1978; Landau & Gleitman, 1985; Miller, 1991).

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 ("goŭ"). The Human Simulation Paradigm (Gillette, Gleitman, Gleitman, & Lederer, 1999) has proved an effective tool for assessing 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. 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 is not quite 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 (Gleitman, Cassidy, Nappa, Papafragou, & Trueswell, 2005;

Pinker, 1984). In the most recent and 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 appears to be the result of joint attention (Baldwin, 1991; Tomasello & Farrar, 1986; Woodward, 2003) and clear visual signatures (Pereira, Smith, & Yu, 2014), especially temporal cues indicative of causation (Scholl & 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 (de Marchena, Eigsti, Worek, Ono, & Snedeker, 2011; Markman, 1990; Markman, Wasow, & Hansen, 2003), especially in the study of syntactic boostrapping (Fisher, Hall, Rakowitz, & Gleitman, 1994; Gillette et al., 1999; Gleitman, 1990; Naigles, 1990).

This paper tries to characterize word learning mechanisms that can make the most out of the rare but highly informative 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 DOG is highly informative as it disambiguates DOG from CAT. 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 PLATE on the kitchen table. Or it may represent a piece of structural information ("a seb" vs. "sebbing") that settles the syntactic category of an unknown word (Brown, 1957). DOG and CAT may also be stand-ins for CHASE and FLEE, PUSH and MOVE, and so on, which are among Lila's favorite verbs (Fisher et al., 1994; Gleitman, 1990). 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).

The resolution of Gleitman's Problem for word learning, then, must make good use of these rare but highly informative nuggets of information. With this in mind, let us review some word learning models on the market.

3. Two worldviews and three learning models

A picture may be worth a thousand words, but what if each word is paired with a thousand pictures themselves? Would the true meanings of words emerge in this statistical aggregation? For example, the word "dog" may, on average, be accompanied with more learning instances that contain DOG than those that contain CAT, and it will thus come out as the winning candidate over time. The modern idea of cross-situational learning (Pinker, 1994; Yu & Smith, 2007) is an update of the associationist program for language (Quine, 1960). In a formulation from that era, "whenever some stimulus other than the significate (e.g., "dog"/DOG; CY) is contiguous with the significate, it will acquire an increment of association with some portion of the total behavior elicited by the significate" (Osgood, Suci, & Tannenbaum, 1957, p. 6). 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 (Le, 2013; Roy et al., 2006), this all seems plausible.

Unless children literally learned everything about a word in one shot—which would result in unrealistically rapid acquisition of the vocabulary—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 (Fazly, Alishahi, & Stevenson, 2010; Frank, Goodman, & Tenenbaum, 2009; McMurray, Horst, & Samuelson, 2012; Siskind, 1996; Yu & Smith, 2007) resolves referential ambiguity by aggregating over word-meaning co-occurrences. The meaning of a word is the candidate with the strongest statistical correlation over all learning instances. By contrast, a *local* approach (Aravind et al., 2018; Medina et al., 2011; Stevens et al., 2017; Trueswell, Medina, Hafri, & Gleitman, 2013) 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 learns from unambiguous data, a powerful idea that has desirable formal properties (Angluin, 1980; Berwick, 1985) and has been proposed for other problems in language acquisition (Fodor, 1998; Wexler & Culicover, 1980).

To understand the differences between the local and global approaches to word learning, consider an illustrative example in Fig. 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 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. Fig. 2a represents the internal state of the model during each learning instance, where each meaning candidate increases its score by one whenever "mipen" is heard. Despite the absence of ELEPHANT in learning instance (d), it still emerges the winning candidate at the conclusion of learning.

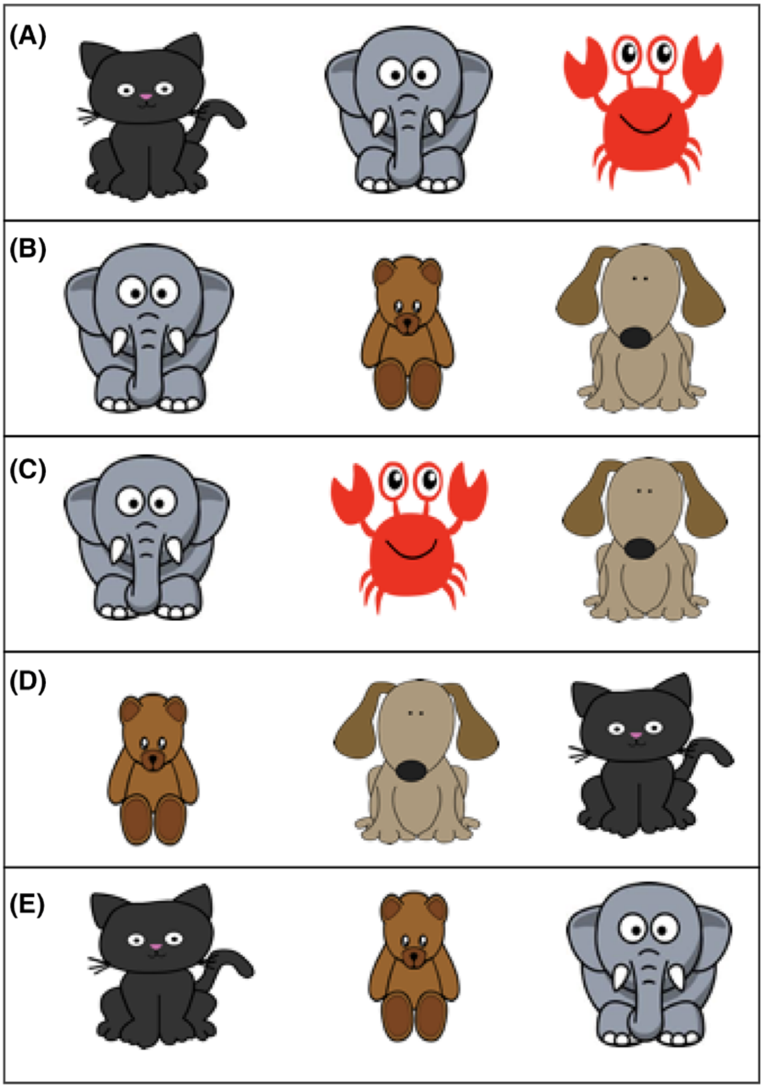
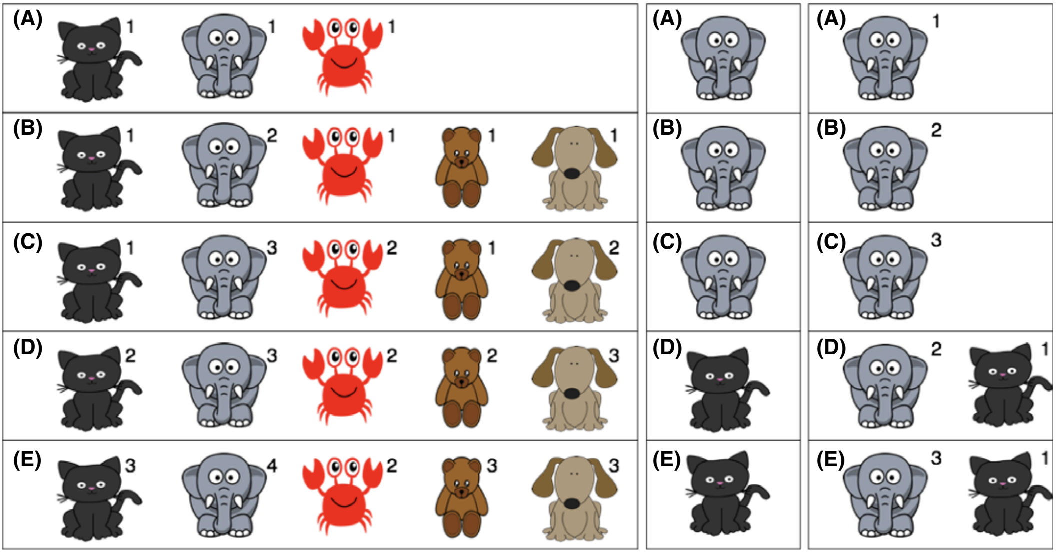


Fig. 1. Mipen. Adapted from Stevens et al. (2017).

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 checks 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 Fig. 2b. For the sake of argument, suppose the learner initially guessed that "mipen" means, correctly, ELEPHANT. 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,



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

Fig. 2. The dynamics of three word learning models: (a) Cross-situational learning, (b) Propose-but-Verify, and (c) Pursuit. Adapted from Stevens et al. (2017).

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 limited size of the internal memory representation (exactly one referent) in Fig. 2b. Note also that the PbV model is stochastic: Because the model "guesses" the intended candidate meaning, it may produce different outcomes on multiple runs through the same input sequence. Thus, Fig. 2b 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. In experimental studies, the PvB model is assessed by aggregating results from multiple subjects, which corresponds to the expected probability of learning the target meaning in the formal analysis presented in Section 4.

The Pursuit model can be viewed as a probabilistic update of PbV. Here, the hypotheses are stored in the memory with weights. Like PbV, Pursuit only memorizes hypotheses considered by the learner but not those that merely co-occur with the word—hence the slightly wider internal state in Fig. 2c than PbV but much smaller than the cross-situational learner which keeps track of everything. 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 & Mosteller, 1951; Dayan & Daw, 2008; Rescorla & Wagner, 1972; Sutton & Barto, 1998) 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 for a word, the learner will single-mindedly *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 Fig. 2c is only one of the possible paths available to the learner. The score keeping of Pursuit is somewhat complex (see Stevens et al., 2017 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 by scene (c). Note that since ELEPHANT has been confirmed, the learner does not register any other referents at all. At instance (d), however, the absence of ELEPHANT lowers its score 2. The learner adds another hypothesis, say CAT, to the list with a score of 1. In 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 meant 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, Trueswell, & Gleitman, 2013).

Stevens et al. (2017) provide extensive empirical and computational analyses of the three models. The results include a surprising finding 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., Köhne et al., 2013; Medina et al., 2011; Trueswell et al., 2013) provide more direct and decisive support for the local models over global models.

Here, I focus on another surprising result from our study, a quantitative evaluation of how these models fare "in the wild." To maintain continuity with previous modeling efforts (Frank et al., 2009; Yu & Ballard, 2007), we annotated two videotaped sessions 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 utterance. There are 34 unique words in all. We implemented the PvB model, the Pursuit model, as well as a cross-situational learning Fazly et al. (2010) that makes learning decisions on the overall tallies of word-meaning co-occurrences as described in Fig. 2a. After the entire corpus is processed, the models' output is compared against a gold standard lexicon. Three performance measures are reported: precision, recall, and F-score. Precision gives the percentage of the correct word-referent pairs learned by the model, recall measures the percentage of the target word-referent pairs in the gold standard lexicon that the model manages to learn, and *F*-score is a composite measure of the two. As is conventional in computational modeling research, the free parameters for

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

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

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. The results are summarized in Table 1.

It must be said that the results in Table 1 are terrible: Parents would be running to the family pediatrician if children were only able to learn more words incorrectly than correctly. At the same time, it is worth noting that the only information these learning models capture is one component of learning—the tracking and computation of word-meaning co-occurrences—and do not incorporate any linguistic or nonlinguistic constraints well documented in the empirical literature. The models' abysmal performance only highlights the necessity of such constraints for successful language acquisition.

The most surprising finding here is the unreasonable effectiveness of Pursuit: A local, resource-limited model that only considers a subset of word-referent pairings actually outperforms the cross-situational model that optimizes over the totality of the learning data. This will be the primary focus of our formal and quantitative analysis in Section 5, but a related, and intriguing, pattern emerged in the analysis of errors produced by these learning models, as shown in Fig. 3.

The graphs in Fig. 3 compare Pursuit and cross-situational learning for "easy" and "hard" words. Here, the difficulty of a word is defined as its average ambiguity across all learning instances. For example, if a visual scene that accompanies the utterance of a word contains only one noticeable object (as judged by the human annotator), the word will have an ambiguity score of 1. If there are six objects, then the word will have an ambiguity score of 6. The distribution of ambiguity across learning instances is broadly consistent with the findings in Trueswell et al. (2016): Most of times a word co-occurs with several objects but occasionally the learning instance is highly salient (e.g., only one referent). Fig. 3 shows that for words with an ambiguity score greater than or equal to 3 ("hard"), both models learn rather poorly, with the cross-situational model making only three correct mappings and the Pursuit model making four correct mappings. But when we look at those words that have an average ambiguity score less than 3 ("easy"), we see that while the cross-situational model does not do much better on these words (four correct mappings), Pursuit learns more than twice the number of words (nine correct mappings). There are 12 out of 34 words in the gold standard lexicon that have the average ambiguity less than 3. Tellingly, three quarters (9/12) of these are learned correctly by the Pursuit model, more than double the accuracy of the cross-situational model (4/12).

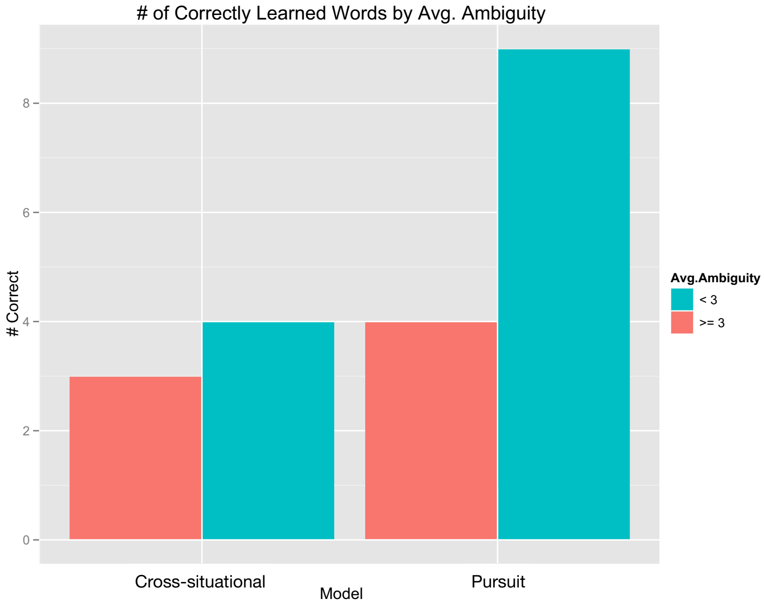


Fig. 3. Performance of Pursuit and cross-situational models on high- and low-ambiguity words.

What makes a resource-limited model more effective than one that retains the totality of information? Why is it especially effective for low-ambiguity easy words? I now turn to investigate these puzzling results.

4. The race between DOG and CAT

Let us revisit DOG versus CAT, a toy example that nevertheless encapsulates the problem of ambiguity resolution in word learning. As reviewed in Section 2, word-meaning associations are generally highly ambiguous, but there are rare instances of clarity. For the present analysis, it is not important what makes these instances informative (see Section 2); it is sufficient to assume that they exist albeit 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 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.

4.1. Formal analysis

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 us assume that the input sequence starts with N learning instances in which both DOG and CAT are present, followed by U—which stands for unambiguous—instances where the competitor CAT is not present (but DOG is). Furthermore, let us assume that in each of the first N instances, DOG and CAT have an equal probability of p_N of being selected. In the U instances that follow, DOG has a probability p_U of being selected each time; CAT will not be chosen because it is not present.

Here, p_N and p_U can be flexibly defined. For instance, they may be inversely proportional to the number of potential meaning candidates in each instance, a frequent manipulation in the experimental study of word learning (Köhne et al., 2013; Medina et al., 2011; Yu & Smith, 2007). For our purposes, we generally assume $p_N < p_U$; that is, the N instances are not as helpful as the U instances for learning DOG. In fact, the Human Simulation Paradigm research reviewed in Section 2 suggests that in general, $U \ll N$ and $p_U \gg p_N$. That is, most learning instances are highly ambiguous, but the learner occasionally receives highly informative cues.

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 will on average be in a tie. We are thus interested in quantifying the effectiveness of the U instances as tie-breakers: The further that the U instances push DOG ahead of CAT, the more likely will the learner succeed.

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):

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

Note again that we are only considering the competition between DOG and CAT, 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, he or she would be at 50% under all learning models. The term advantage is a quantitative measure of the "above chance" boost conferred by the subsequent U disambiguating instances.

Strictly speaking, a "true" cross-situational learner (Fazly et al., 2010; Frank et al., 2009; Yu & Smith, 2007) will always select DOG with probability 1 because DOG will have a higher score than CAT. But the calculation of the cumulative advantage afforded by U in Eq. 1, which is a ratio and thus a gradient measure, is appropriate for at least two reasons. First, experiments (Medina et al., 2011; Smith & Yu, 2008; Trueswell et al., 2013) clearly show that for word learning across multiple instances, the target meanings are only preferred rather than categorically selected; see Stevens et al. (2017) for discussion. This suggests that a cross-situational learning model must have a probabilistic/scalar component that encodes the learner's confidence in DOG over other meaning candidates, such as the one described in Eq. 1. Second, 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 & Schooler, 1991), to determine whether a meaning has been successfully acquired (e.g., Fazly et al., 2010; Frank et al., 2009; Stevens et al., 2017). These threshold functions operate on the *relative* strength of the candidates. If the advantage of DOG 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 the 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 10 distractors. More concretely, let us introduce a "bias" factor for the U informative instances, which is simply taken to be the ratio of p_U and p_N; thus, more informative learning instances are weighed more heavily. The resulting advantage for DOG over CAT for this 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 ignoring 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 selects 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.
- 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 that pushes DOG ahead of CAT. Thus, the advantage for a Pursuit model (Δ_P) afforded by U is the product of two probabilities: that of CAT being ahead of DOG in *N,* and that of DOG catching up in *U*.

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

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*) 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.

4.2. Numerical results

The discussion so far has been abstract. The effectiveness of the models, which is measured by the advantage value Δ, is determined by the values of several parameters that specify the extent of referential ambiguity in word learning. We now explore these parameter values so as to understand the conditions under which the learning models are most effective, and how such conditions correspond to realistic word learning situations. Here, the results from the Human Simulation Paradigm studies (Trueswell et al., 2016) are critical, as they give us a realistic estimate of the ambiguity problem in word learning, thereby providing some baseline values for the parameters.

The first result focuses on how to make effective use of rare but highly informative learning instances. Following Trueswell et al. (2016) (briefly reviewed in Section 3), I will assume $P_N = 0.2$, which is roughly the average/baseline 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. Fig. 4 shows the relative advantage of DOG over CAT (Δ) as a function of *U,* the number of disambiguating learning instances, which take place after $N = 100$ ambiguous instances.

It is clear from Fig. 4 that the advantage of DOG over CAT is significantly stronger under Pursuit than the cross-situational learning models, as long as the informative learning instances are relatively rare (the value of *U* being small). Note the effectiveness of Pursuit is not absolute. For instance, when the value of *U* is large, that is, when the informative instances are common, the biased cross-situational model begins to outperform the Pursuit model although the vanilla cross-situational model continues to struggle. In the present context, however, it is important that Pursuit excels precisely under the conditions that are most similar to actual word learning situations, namely, highly ambiguous input with relatively few informative instances for the purpose of disambiguation.

Fig. 5 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): A larger value indicates the higher informativeness of *U* (over

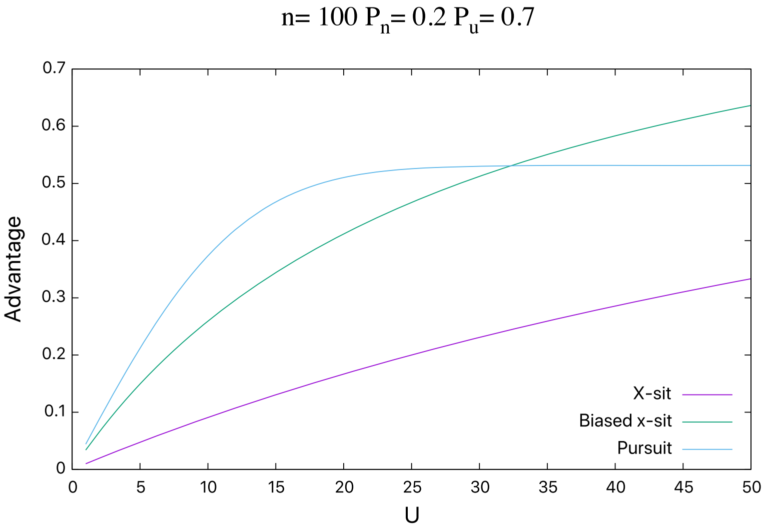


Fig. 4. Pursuit makes better use of rare but informative learning instances.

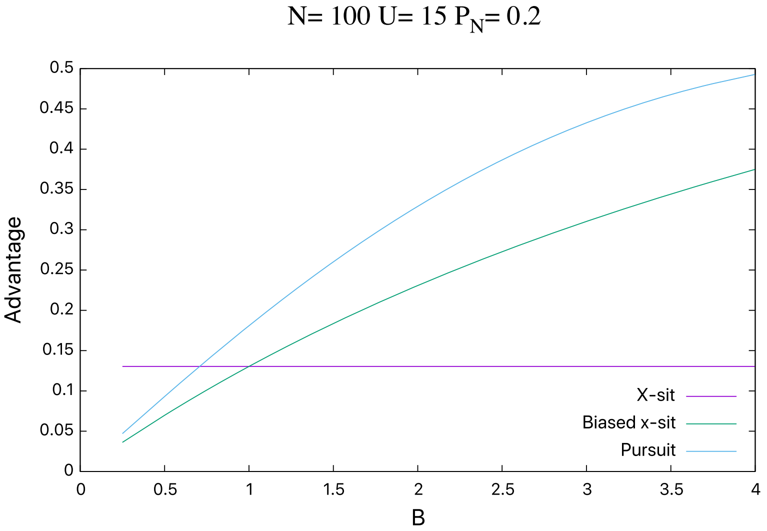


Fig. 5. Pursuit makes better use of more informative learning instances.

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 vanilla cross-situational model, which only tallies the number of word-referent occurrences, is not able to take advantage of the more informative learning instances. The effectiveness of Pursuit over the biased cross-situational learning model, however, is decisive. This is even so when $B < 1$; that is, the U instances are more ambiguous than the *N* instances. Nevertheless, DOG is able to receive a considerable advantage over CAT, which is in fact absent.

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 Figs. 4 and 5 suggest that a Pursuit-like model is better equipped than cross-situational models to capitalize on 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 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. This, I believe, is the reason why the Pursuit model performs much better than cross-situational learning for the lower ambiguity words as shown in Fig. 3: Low-ambiguity words correspond to higher values of *B* in Fig. 5, where Pursuit holds a significant advantage.

5. Summary

This schematic paper seeks to identify word learning algorithms that can make the most out of very limited informative data in the environment, and to explain some puzzling results from the word learning literature. The Pursuit model, which only attends to a very limited subset of the learning data, may also have some interesting developmental implications. It has long been conjectured that children's cognitive resource limitations may turn out to be a blessing in disguise, enabling them to focus on the critical aspects of the learning task at hand (Elman, 1993; Newport, 1990; see Yang, 2016 for a learning-theoretic underpinning of this idea). Perhaps a young child—or even adults as in the many behavioral experiments that motivated the Pursuit model—do not have the cognitive capacity to entertain more than a couple of hypotheses for the meanings of words, and this turns out to just the right approach to Gleitman's Problem in the messy real world. The signal does exist, but it is faint; the worst-case scenario would be for it to be drowned out by a "big data" model that tabulates and deliberates over everything. As Lila has taught us, word learning is hard; perhaps the best solution is to not try harder.

Acknowledgments

I would like to thank my coconspirators Lila Gleitman, John Trueswell, and Jon Stevens for part of the research presented here. The comments from Barbara Landau and two anonymous reviewers have also been very helpful.

Notes

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

2. Most reinforcement learning models sample among the hypotheses according to their probabilities. If so implemented, the probabilistic model becomes formally equivalent to an online version of cross-situational learning and will inherit all the empirical problems pointed by Stevens et al. (2017). 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 & 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 (Lignos, 2013; Murray & Forster, 2004; Sternberg, 1969; Yang, 2016), which corresponds to the number of confirmations in the Pursuit model.

References

Anderson, J. R., & 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.

Aravind, A., de Villiers, J., Pace, A., Valentine, H., Golinkoff, R., Hirsch-Pasek, K., Iglesias, A., & Wilson, M. S. (2018). Fast mapping word meanings across trials: Young children forget all but their first guess. *Cognition*, *177*, 177–188.

Baldwin, D. A. (1991). Infants' contribution to the achievement of joint reference. *Child Development*, *62*, 875.

Berwick, R. (1985). *The acquisition of syntactic knowledge*. Cambridge, MA: MIT Press.

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

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

Carey, S., & Bartlett, E. (1978). Acquire a single new word. *Child Language Development*, *15*, 17–29.

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

Dayan, P., & 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., & Snedeker, J. (2011). Mutual exclusivity in autism spectrum disorders: Testing the pragmatic hypothesis. *Cognition*, *119*(1), 96–113.

Elman, J. L. (1993). Learning and development in neural networks: The importance of starting small. *Cognition*, *48*(1), 71–99.

Fazly, A., Alishahi, A., & 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., & 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., & 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 S. A. Kuczaj (Ed.), *Language, thought, and culture* (pp. 301–334). Hillsdale, NJ: Lawrence Erlbaum.

Gillette, J., Gleitman, H., Gleitman, L., & 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., & Trueswell, J. C. (2005). Hard words. *Language Learning and Development*, *1*(1), 23–64.

Gleitman, L. R., & 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., & 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., & Gleitman, L. R. (2013). Multiple proposal memory in observational word learning. In M. Knauff et al. (Eds.), *Proceedings of the 35th Annual meeting of the Cognitive Science Society* (pp. 805–810). Austin, TX: Cognitive Science Society.

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

Le, Q. V. (2013). Building high-level features using large scale unsupervised learning. In J. Langford & J. Pineau (Eds.), *Acoustics, speech and signal processing (ICASSP), 2013 IEEE International Conference* (pp. 8595–8598). IEEE.

Levin, B. (1993). *English verb classes and alternations: A preliminary investigation*. Chicago: University of Chicago Press.

Lignos, C. (2013). Modeling words in the mind. PhD thesis, University of Pennsylvania.

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

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

Markman, E. M., Wasow, J. L., & 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., & 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., & 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*. San Francisco, CA: Scientific American Library.

Murray, W. S., & 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., & Tannenbaum, P. H. (1957). *The measurement of meaning*. Urbana: University of Illinois Press.

Pereira, A. F., Smith, L. B., & 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*. Cambridge, MA: Harvard University Press.

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*. Cambridge, MA: MIT Press.

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*. Vol. *2* (pp. 64–99). New York: Appleton Century Crofts.

Roy, D., Patel, R., DeCamp, P., Kubat, R., Fleischman, M., Roy, B., Mavridis, N., Tellex, S., Salata, A., Guinness, J., Levit, M., & Gorniak, P. (2006). The human speechome project. In P. Vogt et al. (Eds.), *Lecture notes in computer science* (pp. 192–196). New York: Springer.

Scholl, B. J., & 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-tomeaning mappings. *Cognition*, *61*(1), 39–91.

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

Sternberg, S. (1969). Memory-scanning: Mental processes revealed by reaction-time experiments. *American Scientist*, *57*(4), 421–457.

Stevens, J. S., Gleitman, L. R., Trueswell, J. C., & Yang, C. (2017). The pursuit of word meanings. *Cognitive Science*, *41*, 638–676.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. New York: 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., & Farrar, M. J. (1986). Joint attention and early language. *Child Development*, *57*, 1454–1463.

Trueswell, J. C., Lin, Y., Armstrong, B., Cartmill, E. A., Goldin-Meadow, S., & 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., & Gleitman, L. R. (2013). Propose but verify: Fast mapping meets cross-situational word learning. *Cognitive Psychology*, *66*(1), 126–156.

Wexler, K., & Culicover, P. (1980). *Formal principles of language acquisition*. Cambridge, MA: MIT Press.

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, UK: Oxford University Press.

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*. Cambridge, MA: MIT Press.

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

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