How words can and cannot be learned by observation

Tamara Nicol Medina[a,b,1], Jesse Snedeker[c,1], John C. Trueswell[a,b,1], and Lila R. Gleitman[a,b,1]

[a]Department of Psychology and [b]Institute for Research in Cognitive Science, University of Pennsylvania, Philadelphia, PA 19104; and [c]Department of Psychology, Harvard University, Cambridge, MA 02138

Contributed by Lila R. Gleitman, April 2, 2011 (sent for review November 23, 2010)

Three experiments explored how words are learned from hearing them across contexts. Adults watched 40-s videotaped vignettes of parents uttering target words (in sentences) to their infants. Videos were muted except for a beep or nonsense word inserted where each "mystery word" was uttered. Participants were to identify the word. Exp. 1 demonstrated that most (90%) of these natural learning instances are quite uninformative, whereas a small minority (7%) are highly informative, as indexed by participants' identification accuracy. Preschoolers showed similar information sensitivity in a shorter experimental version. Two further experiments explored how cross-situational information helps, by manipulating the serial ordering of highly informative vignettes in five contexts. Response patterns revealed a learning procedure in which only a single meaning is hypothesized and retained across learning instances, unless disconfirmed. Neither alternative hypothesized meanings nor details of past learning situations were retained. These findings challenge current models of cross-situational learning which assert that multiple meaning hypotheses are stored and cross-tabulated via statistical procedures. Learners appear to use a one-trial "fast-mapping" procedure, even under conditions of referential uncertainty.

acquisition | induction | language | vocabulary

Fundamental for each child entering the human community is the acquisition of word meanings: discovering which language sounds map onto which interpretations. Because these mappings are arbitrary and vary cross-linguistically, growing a vocabulary poses a classic learning problem for humans, both infant learners of a first language and second-language learners who must replace the original mappings with a new set. This experience-dependent learning problem for humans contrasts with animal communication systems in which the interpretations of species-specific barks, chirps, and growls are largely given for free by nature. The present article provides experimental evidence concerning the primitive initial procedure by which humans acquire vocabulary items.

A common assumption is that form-to-meaning mappings are discovered in a process mediated by observation of extralinguistic events: *The learner matches recurrent speech events to recurrent aspects of the observed world*. For example, when an English speaker says "dog" or a French speaker says "chien," there is likely to be a co-occurring dog sighting. Young children often acquire a word's meaning after a single such exposure to its use in context (1), particularly if there is strong pragmatic support (2) or a restrictive syntactic environment (3). The sheer size of the average vocabulary at age 6 y [estimated at 6,000–8,000 words (1)] suggests that this "fast mapping" of a sound segment onto its interpretation must happen very often, as is also attested in many laboratory studies (4, 5).

Yet the world of words and their contexts is enormously complex. Few words are taught systematically, even in middle-class environments with their blocks and picture books. Rather, in most instances, the situations of word use arise adventitiously as adults interact socially with novices. Words are heard buried inside multiword utterances and in situations that vary in almost endless ways—the bath, the zoo, the supermarket—so that usually a listener could not be warranted in selecting a unique interpretation for a new item. For example, in the fairly typical setting of Fig. 1*A*, there are hundreds of objects, happenings, properties, and relations that might be picked out as a match when the adult utters some new word: one of the objects (shoe, pacifier), qualities (black, large) or parts (lace, sole) of the object, the child's own actions (bump, look), and so forth. Worse, the match could be to any two or more of these categories, e.g., "black shoe" or "Mommy's shoe." Also, speakers often allude to absent objects and events ("Let's go to the *zoo*," "Remember when we lost your *ball*?") or ineffable properties ("Be *nice* to your little brother," "*Time* for your nap."). A word once heard may not be encountered again for days or weeks, during which hundreds of other words intervene. At the other extreme, those things most consistently present—the ceiling, an eye, breathing—are often least likely to be topics of conversation; rather, their omnipresence diminishes their salience as conversational focus. Thus, to say that a word is learned "by" observation must be true in some way, but so saying leaves much to the imagination about actual procedures.

A venerable assumption that addresses some of these problems is that learning happens over several encounters: *Word meaning can be acquired probabilistically and incrementally by accumulating evidence across several situations in which the same phonological segment occurs*. Hearing the same sound under new circumstances can eliminate false conjectures or add detail to underspecified conjectures. Many researchers have tested this hypothesis with laboratory simulations (6–10).* In these experiments, pictures of objects appear on a neutral background (Fig. 1*B*). Each image is assigned a nonsense word, with the participants' task being to discover these mappings. To reproduce uncertainty, several images are shown simultaneously on each trial, along with an equal number of auditorily presented nonsense words. Over trials, participants could track co-occurrences to determine which sound-image pairs are reliably copresent. Adults and children are quite adept at solving such laboratory problems, performing above chance after several such exposures.

At least two major questions, discussed below, arise about the relevance of such laboratory demonstrations to real vocabulary acquisition.

Contextual Uncertainty. As already suggested, in ordinary conversational settings hundreds of objects are in view, rather than three or four, many of them plausible referents. A theory stipulating that all such context-appropriate alternatives are stored, awaiting cross-situational disambiguation, grows less appealing as their number increases. That is, constraints on memory for past contexts may be a more severely limiting factor in natural learning than appears in stripped down laboratory demonstrations. Far more troublesomely, simulating "context" using

Author contributions: T.N.M., J.S., J.C.T., and L.R.G. designed research; T.N.M. and J.S. performed research; T.N.M., J.S., J.C.T., and L.R.G. analyzed data; and T.N.M., J.S., J.C.T., and L.R.G. wrote the paper.

The authors declare no conflict of interest.

[1]To whom correspondence may be addressed. E-mail: gleitman@psych.upenn.edu, medinatn@sas.upenn.edu, trueswell@psych.upenn.edu, or snedeker@wjh.harvard.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1105040108/-/DCSupplemental.

*The kinds of words studied here (basic-level whole-object terms that surface as nouns in most languages) constitute only a small proportion of words known and used by 3- and 4-y-olds. Nevertheless, these items predominate in the vocabularies of infants (11) and novice older learners of both first and second languages (12). These words constitute the enabling basis for more advanced vocabulary growth (1, 3) and require the least internal linguistic support to acquire (13, 14). Abstract words such as *time* or *think* have no identifiable images and are acquired via procedures that make crucial use of their syntactic and discourse contexts (15).

Fig. 1. (*A*) A plausible word learning environment for the word shoe. (*B*) The simulated word-learning environment for shoe found in most cross-situational word-learning experiments.

repeated identical images seriously distorts reality, for learners hear the word *dog* not always and only in the presence of some dog (or dog-image) frozen in time and place, but in the presence of different dogs doing different things—and sometimes in the absence of any dog . Hence, even with perfect memory, one may never settle on a single correct meaning as exposures accumulate. Because words are uttered from time to time absent their referents, the number of available hypotheses could actually increase across contexts, posing a principled conundrum for statistical learning machinery. A collateral possibility is that learners may have implicit means for distinguishing between more and less useful contexts, discarding some input without its entering into the search for meaning (16). Our first aim in the present experiments will be to understand the extent to which statistical-accrual procedures, demonstrable for small fixed sets of word-image pairings, scale up to the multiply ambiguous contexts in which vocabulary is acquired.

Evaluating the Models. Although the models usually proposed for laboratory word-learning findings are information-accrual based, their major computational assumptions have not been tested behaviorally, but simply taken for granted. In outline, these schemes propose that for each learning instance, all concepts known to the learner are cross-tabulated with the word. If the situation is interpreted as supporting certain pairings, those associations are strengthened (e.g., hearing "moop" in the presence of Fig. 1*B* leads to simultaneous strengthening of the *moop*-shoe and *moop*-chair associations, among others). Later encounters with *moop* in other contexts lead to additional associative updates. As this theory implies, even when each learning instance is low in informativity (because any of the objects on the screen could be the referent of *moop*), learners gradually reach consensus by tracking and retaining all observed pairings, with the strongest associate chosen at the end. However, no time-course data are provided for learning across instances that might validate these assumptions. Rather, the only measure reported is final response accuracy (above-chance performance after several image-to-sound pairings), an outcome that (depending on the dataset) might be generated by all-or-none models as well as statistical-accrual ones. [One study (10) did collect trial-by-trial conjectures, but did not consider the possible strategy of remembering a single hypothesis across trials.] In the experiments reported below we also examine how learners' interpretations evolve across successive contextually uncertain encounters.

Experimental Findings

In three experiments, participants guessed word meanings by observing 40-s videos ("vignettes") of naturally occurring contexts (the meal, the play-room, and so forth) in which parents uttered these words to their infants aged 12 to 15 mo. For example, one vignette showed a mother opening a bag of toys while saying, "Who else is in the bag?" to her child. Participants watched the vignettes with the audio muted. They heard only an identifying signal (either a beeping noise or a nonsense word, depending on the experiment) occurring exactly when the mother had (really) said, e.g., "bag." Because new words are rarely introduced in an uninterrupted sequence ("doggie, doggie, doggie. . .") but rather intermixed with other words, the vignettes for each target word were distributed among all of the others during the experiments. Given that these vignettes reflected a random sample of parent–child interactions, the vignettes and their presentation conditions reproduce critical properties of real environments in which early word learning happens.

The vignettes were muted for two reasons. First, this replicates the conditions of the experiments just reviewed, which present word-image pairs without sentence context. Second, infants learning their first words do not know the meanings of other words in the utterance and thus cannot use these to determine meaning; by removing this information we more accurately model the earliest stages of acquisition, the point at which statistical accrual is argued to play its greatest role (3, 11). The experiments used these vignette "inputs" to probe the learning procedure for early vocabulary systematically, in terms of three questions: (*i*) When do naturally occurring scenes offer information precise or relevant enough to trigger successful word learning? (*ii*) How do repeated exposures to the same word in different contexts channel and enhance this process? (*iii*) How does memory influence operation of the learning machinery?

Exp. 1: Vignette Informativity. The contexts in which words are uttered vary in informativity, ranging from cases where a parent may gaze fixedly at an object the infant is holding as they name it ("This is a horse") to cases where no horse is even in sight ("Let's go see the horses at the farm"). Here we estimated the frequency of such more and less informative contexts.

Thirty-seven adults watched randomly selected vignettes containing the 24 most frequently occurring nouns and 24 most frequently occurring verbs in the video corpus developed by the author J.S. There were 288 vignettes total, and groups of participants viewed different subsets: two vignettes for each of 48 words. The occurrence of each such "mystery word" was identified by a beep that was identical within and across items, so there was no opportunity for cross-situational learning. The question was which of these vignettes were sufficiently transparent that observers could guess the speaker's beeped-out word. A shorter version of this experiment was run with 12 3- to 5-y-olds, who watched one video for each of eight target words (*SI Text*).

Consistent with our prior findings (14), participants were poor at guessing the intended referent of even highly frequent terms when shown a single instance. Only for 7% of these vignettes did adult participants achieve accuracy scores at or above 50% correct; we classified these vignettes as High Informative (HI). In contrast, 90% of vignettes resulted in accuracy scores below 33%, thus categorized as Low Informative (LI). It is of some interest that none of the obtained HI vignettes were for verbs; all of them were for common nouns labeling categories of whole objects (e.g., *bag, ball, horse*) that are very frequent in the speech of parents to children and appear in the vocabularies of 1- and 2-y-olds world-

wide (11). This pattern is consistent with previous demonstrations (14) that this concrete (or "imageable") component of the lexical stock is the easiest kind to acquire from passive observation of situational context. Sixteen of the 24 nouns examined were of this type; when we restrict the pool of vignettes to these words (i.e., excluding verbs and abstract nouns), the ratio of HI to LI vignettes is approximately 1:5.

We also assessed whether preschoolers were sensitive to the same aspects of the observed events as adults, and they were: Children were more accurate on HI vignettes (53%) than LI vignettes (22%).

In sum, this experiment exhibited the difficulty of determining the meaning of words based on a random sample of single observations, even when items are all highly frequent in parental speech. Expectation of this uncertainty has always motivated cross-situational approaches to word learning. In addition to providing some documentation of the magnitude of uncertainty in parent–child interactive settings, the results provided a set of vignettes that was classified for informativity. These vignettes were used in Exps. 2 and 3 to see how cross-situational opportunities might aid learning. Importantly, Exp. 1 demonstrates that adults and young children cull these vignettes for information in much the same way, so that we could continue to use adults as foils for our learning questions with some confidence that the results would link plausibly to the case of child word learning. These findings further document the accumulating evidence that child and adult word learning share essential commonalities (7, 12, 15).

Exp. 2: Learning with Cross-Situational Information. We now asked how providing several contexts for new words affects learning. We selected concrete whole-object items and set the ratio of HI to LI vignettes at 1:4, to approximate their actual rate of appearance in mother-infant speech as ascertained from Exp. 1. New participants identified 12 such "mystery words" by observing five contexts for each. Each word was assigned a nonsense version (e.g., *vash*), replacing the beeps of Exp. 1 so that participants could determine which vignettes went with which item, even though they were intermixed in presentation order (e.g., a *mipen* vignette, followed by a *vash* vignette, and so forth). Presence and position of HI and LI vignettes was manipulated for four participant groups. One group saw five LI vignettes for each mystery word (HI Absent). The other three groups saw four LI vignettes and one HI vignette, but each in a different order: in one group the HI vignette was the first for each target word (HI First), for another it was third (HI Middle), and for the final group it was fifth (HI Last). Participants recorded their guess after viewing each vignette, allowing for later assessment of how knowledge was evolving across interim information states. They also provided a confidence rating and a final guess for each word at the end of the experiment.

Current statistical models evaluate each word-meaning hypothesis based on the properties of all past learning instances, regardless of the order in which they are encountered (6–10), holding numerous hypotheses in mind (with changing weights) until some learning threshold is reached. This process predicts that: (*i*) final conjectures will be indifferent to the sequential position of HI vignettes; (*ii*) multiple meanings (i.e., traces of past encounters) will be stored and tracked across vignettes; and (*iii*) gradual learning will occur across LI vignettes. Here we test a radically different, one-trial or "fast mapping" hypothesis, in which (*i*) learners hypothesize a single meaning based on their first encounter with a word; (*ii*) learners neither weight nor even store back-up alternative meanings; and (*iii*) on later encounters, learners attempt to retrieve this hypothesis from memory and test it against a new context, updating it only if it is disconfirmed. Thus, they do not accrue a "best" final hypothesis by comparing multiple episodic memories of prior contexts or multiple semantic hypotheses.

The contrasting predictions of these two approaches are clear: if learning is statistical, the position of HI in the sequence of vignette presentations (the three experimental group conditions) will produce perturbations in the learning curve early on but any such effect should be wiped out in the end, for by the fifth vignette all participants in all experimental conditions will have received all and only the same contexts and can compare among them; multiple hypotheses should be present, especially on early vignettes; and performance accuracy should roughly rise across vignettes. However, if learners posit a single hypothesis (and no alternatives) on the first encounter and then seek to confirm or disconfirm it, accuracy on the first vignette should be a major determinant of accuracy by the fifth vignette, there should be little or no evidence in interim vignettes of maintaining back-up hypotheses, and response accuracy need not rise across vignettes.

Results

Effect of Information Order. Fig. 2 plots results as a function of presence and order of HI information. As predicted by the one-trial learning hypothesis, when participants received an informative first exposure (HI First) they tended to: guess correctly (66% accuracy on vignette 1); stick to that guess across vignettes (41% accuracy by the fifth vignette); offer this guess as "the meaning of the word" at experiment's end (37% accuracy on the Final Guess); and record strong and rising cross-trial confidence in this answer (*SI Text*). Accuracy on the fifth vignette and final conjectures was significantly higher in the HI First condition than in third (HI Middle) or fifth (HI Last) conditions (Table 1a, Effect of learning condition on the fifth vignette, and Table 1b, Effect of learning condition on the final conjecture). Not only did good information arriving late fail to lead to correct performance in the end (contra statistical learning predictions), but a set of low-informative vignettes failed to lead to any cross-trial improvement (HI Absent) (Table 1c, Effect of vignette number on accuracy in HI Absent condition).

The consequences of initial correct guessing are most apparent in Fig. 3, which plots accuracy independent of whether input was "good" (i.e., HI) or not. In either case, a major determinant of accuracy on later vignettes is accuracy on the first. [Data from vignette 5 and final conjectures of the HI Absent condition are excluded in Fig. 3 because vignette 5 repeated vignette 1 (*SI Text*).] Symmetrically, performance is extremely poor after an incorrect guess, even when that incorrect guess was made on a HI vignette. This latter finding contradicts current cross-situational models, because it shows that participants had little to no memory of alternative word-meaning hypotheses that they could return to. This finding vindicates the reality of the issue mentioned in introductory remarks, namely the impossibility of storing the scores (if not thousands) of objects, qualities, and so forth, that any single observation embodies, and that might be relevant to the meaning of a new word then heard. Had participants stored some plausible alternative hypotheses after guessing incorrectly on a HI vignette, one of the hypotheses would likely have been the correct one, leading to better performance

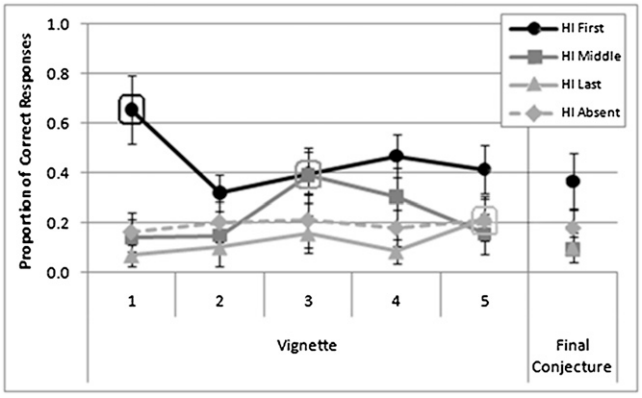

Fig. 2. Mean proportion of correct responses for each learning instance (vignette) and the final conjecture. Error bars = 95% confidence intervals, Exp. 2.

Table 1. Exps. 2 and 3: Results of multilevel logistic regressions of binary data

Model	Effect	Estimate	SE	Wald-Z	*P* value
a. Effect of learning condition on	Intercept	−0.5941	0.5719	−1.039	0.29893
accuracy of the fifth (Final) vignette.	HI First vs. HI Middle	−1.9654	0.4935	−3.983	0.00007***
(Baseline = HI First)	Hi First vs. HI Last	−1.2564	0.4099	−3.065	0.00218**
b. Effect of learning condition on	Intercept	−0.7621	0.4724	−1.613	0.10700
accuracy of final conjecture.	HI First vs. HI Middle	−1.8978	0.4571	−4.152	0.00003***
(Baseline = HI First)	HI First vs. HI Last	−1.963	0.4902	−4.004	0.00006***
c. Effect of vignette number (1–5) on	Intercept	−2.7558	0.86297	−3.193	0.00141**
accuracy in HI Absent condition.	Vignette	0.02487	0.08608	0.289	0.77263
d. Effect of first vignette accuracy on	Intercept	−2.9295	0.6467	−4.530	< 0.00001***
accuracy of vignette 2.	Accuracy of V1	2.0704	0.2952	7.013	< 0.00001***
e. Effects of confirmation and accuracy	Intercept	−2.7684	0.222	−12.471	< 0.00001***
on repetition of response	Confirmation	1.9742	0.2298	8.59	< 0.00001***
	Accuracy	2.0269	0.2414	8.397	< 0.00001***
f. Effect of delay on accuracy of the HI	Intercept	−1.5874	0.4281	−3.708	0.00021***
vignette. HI Last.	Delay	1.3611	0.4892	2.782	0.0054**

Models computed in R, using crossed random effects for Subjects and Items. See *SI Text* for R-syntax. Except for the model, Effect of vignette number (1–5) on accuracy in HI Absent condition, models were significantly better fits compared with simpler models which did not include the reported fixed effects [χ^2 tests of change in −2 restricted log likelihood (20)]. More complex models with additional fixed effects or interactions did not reliably improve fit. The "Effect of vignette number (1–5) on accuracy in HI Absent condition" model is shown for illustration, as it is *not* a better fit than the null model. In the "Effect of first vignette accuracy on accuracy of vignette 2" model, models including Informativeness of the first vignette (HI vs. LI) did not improve the fit. ***$P < 0.001$, **$P < 0.01$, *$P < 0.05$.

later compared with initially guessing incorrectly on a LI vignette (Table 1d, Effect of first vignette accuracy on accuracy of vignette 2). Statistical accrual models, (e.g., ref. 6), store all plausible hypothesized meanings for later consideration on the next learning instance.

Confirmation effect. A single-conjecture learning procedure based on variable input would seem likely to acquire a vocabulary replete with errors, with one child calling the shoes "shoes" but another child calling them "wheels" or "dogs," depending on the vagaries of first encounters. However, such errors are vanishingly rare, even in toddlers. If learners are not (as in associative schemes) maintaining and cross-tabulating several hypotheses, what is preventing such errors? A clue comes from a further feature of Fig. 2. Rather than accuracy rising between vignettes 1 and 2, as would be expected if several hypotheses were being evaluated across learning instances, accuracy actually drops, a so far unexplained and surprising consequence of added contextual information. This decrement over vignettes is true even in our "best" case of HI data received on vignette 1, for which the accuracy of over 60% on vignette 1 falls to about 40% by vignette 5.

Part of this performance drop no doubt reflects participants' failing on occasion to recall their initial hypothesis, for overall the experiment imposes a considerable memory burden. Even

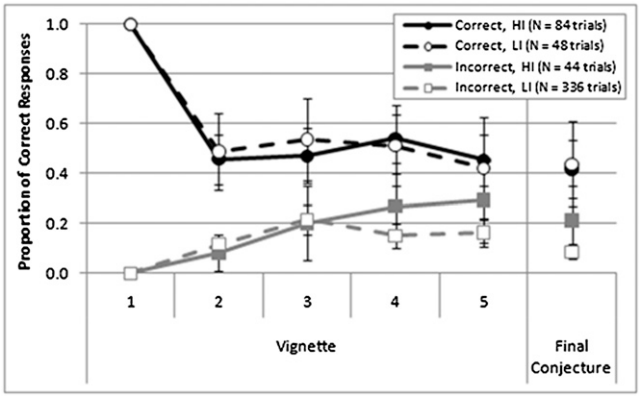

Fig. 3. Mean proportion of correct responses for each vignette and the final conjecture, plotted as a function of being correct or incorrect on vignette 1, split by whether vignette 1 was High informative (HI) or low Informative (LI). Error bars = 95% confidence intervals, Exp. 2.

closer to the heart of the learning problem for vocabulary is participants' likely bemusement by the striking differences in contexts across vignettes, an issue that cannot arise in repeated image-to-label experimental studies. Overall, the rarity of context-invariance suggests an implicit reliability check on the initial conjecture before it is consolidated in memory.

Exp. 1 results provided the basis for investigating this issue. Participants in Exp. 1 provided a range of conjectures for each vignette, viewed in isolation. We reasoned that subsequent vignettes in Exp. 2 would provide confirming evidence for the initial hypothesis (the guess on vignette 1) just in case that hypothesis was a member of the Exp. 1's response set for that word. That is, if a vignette in isolation brings to mind a shoe, even if ever so slightly, enough to have drawn a *shoe* response from at least one of the 12 Exp. 1 participants who observed this item, this would be enough to confirm an initial hypothesis of *shoe* in the cross-situationally exposed participants. Indeed, this is the case. The rate at which a participant repeated a response on the next learning instance was reliably higher when the next vignette confirmed that previous response (0.45) than when it did not (0.19): confirmation quadruples the odds of response repetition (Table 1e, Effects of confirmation and accuracy on repetition of response). This finding is true for both correct (0.53 vs. 0.37) and incorrect guesses (0.37 vs. 0.08), although getting confirmation of an incorrect guess is exceedingly rare (only 133 of 1,536 incorrect responses were confirmed by our criterion).

Taken together, these findings suggest that participants seek supportive evidence in later encounters with a word; for correct conjectures, confirmation is usually found. Thus there is a one-and-a-bit learning procedure with a built-in mechanism for looking before you leap to a permanent entry in the mental lexicon. Finally (see again the predictions above for one-trial learning, and Fig. 2), note that late (HI Middle) recipients of good information do not maintain it, perhaps because they are in the grip of their first, false, hypothesis.

In sum, Exp. 2 findings document a learning procedure in which learners form a single conjecture, retaining neither the context that engendered it nor alternative meaning hypotheses. The criterion for retention of this conjecture is whether it is even minimally plausible for the context in which one next hears the word. A corollary is that a false hypothesis, once formed, blocks the formation of new ones. A by-product of this first-is-best (or "template") procedure is disproportionate dependence on the presence of useful information first. For this reason, receiving a LI vignette first typically results in participants never settling on any one meaning for that word (we will revise the term "never"

when we introduce Exp. 3). This learning procedure of conjecture coupled with verification is most clearly revealed by considering what happens across successive vignettes as a function of HI position. Participants in the HI First condition have the advantage of being more likely to start out with a correct hypothesis just because the first vignette they encounter is a highly informative one, and therefore they will be more likely to find confirming evidence in later learning instances compared with those who began with an incorrect hypothesis. Almost by definition, later vignettes contain information that is more likely to serve as confirmation of a correct guess than an incorrect one (e.g., a later *shoe*/*vash* vignette is more likely to be an event that involves another shoe than another elephant).

Exp. 3: "Never" Isn't "Forever." Results thus far suggest that word learning requires useful information to be present on a first encounter. But we know that so unforgiving a procedure can't describe reality; failing to acquire *cow* the first time you heard it cannot doom you to a lifelong *cow*-gap in your mental lexicon. So there must be a way to interpret the notion of "first encounter" that does not have this fatal implication. To explore this issue, we reran the HI Last and HI Middle conditions of Exp. 2, except now participants had a 1- to 3-d break before encountering their first HI vignette. Such a delay might prevent interference from earlier LI vignettes via ordinary processes of forgetting.

Results

Indeed, accuracy on a HI vignette in fifth/final position (HI Last) was significantly improved when it appeared first on a new day (0.47) compared with performance in Exp. 2 (0.22) ($P < 0.01$) (Table 1f, Effect of delay on accuracy of the HI vignette HI Last). Similarly, accuracy on a HI Middle vignette improved numerically (0.45 vs. 0.39), as did final conjectures for both (HI Last: 0.18 vs. 0.10; HI Middle: 0.17 vs. 0.09), although these latter differences were not significant. This pattern suggests the delay between observations allowed participants to begin learning anew. Whatever false conjectures were entertained in the first experimental session were apparently forgotten, reducing interference in the second session. Performance did still drop somewhat on final conjectures, as in Exp. 2. Nevertheless, a long delay after poor learning instances precludes false conjectures to a useful degree. In effect, "HI Middle" and "HI Last" have been converted to a new "HI First" by the passage of time. So tomorrow is another day for one-trial word learning, just as it is for many other quotidian events.

Discussion

All attempts to understand vocabulary learning incorporate a component that matches recurrent sounds with their situational contingencies. Although it is now recognized that concomitant linguistic and discourse information contributes heavily to word learning (3), the observational component predominates at the earliest stages and plays a continuing role throughout life. Because contexts are so various and uncertain, it has generally been believed that learners store several of them, at some point retrieving and comparing among them to extract their recurrent properties from among the variable ones. This compare-and-contrast supposition lies at the heart of recent cross-situational learning experiments that attempt to simulate novice vocabulary learning and model its character. The present findings challenge this picture. Specifically, our results are incompatible with three fundamental predictions of cross-situational learning: (*i*) indifference to input order, (*ii*) maintenance of multiple hypotheses, and (*iii*) improvement across trials.

Indifference to Input Order. If the learning procedure accrues information across exposures, all of it available for evaluating the best fit at the end, it should not matter when in the input sequence the best information occurs. However, our results show a massive effect of input order, with the first exposure being decisive (and therefore performance being best in the HI First condition), and essentially no learning after a false initial guess in any condition (Figs. 2 and 3).

Maintenance of Multiple Hypotheses. For the proposed learning machinery to benefit from successive exposures, it must maintain several options during early points in the input sequence. However, our laboratory learners' characteristic style was to store a single hypothesis (Table 1d, Effect of first vignette accuracy on accuracy of vignette 2), abandoning it only when disconfirmed (Table 1e, Effects of confirmation and accuracy on repetition of response).

Improvement Across Trials. Another hallmark of cumulative learning is increasing accuracy (though with perturbations) across trials. However, in our studies, accuracy remains flat or falls across vignette trials (Fig. 2). Only on a new day, when the slate is wiped clean, does performance rise again (Table 1f, Effect of delay on accuracy of the HI vignette HI Last).

Our findings, to the extent they can be linked to learning in vivo, begin to suggest that observational word learning is hardly "cross-situational" in the intended sense but is materially an insightful and almost indefeasible process. This outcome is consistent with several recent laboratory demonstrations of young children acquiring new words from single encounters, and then seamlessly generalizing their use (17); it is in accord, as well, with the astonishing rate of child vocabulary growth.

We believe that statistical-accrual approaches to word learning have seemed plausible to many commentators, in part because the learner's task has been informally envisaged as selecting among a few alternative interpretations, more as depicted in Fig. 1*B* than as in Fig. 1*A*. Indeed, statistical models have proven adequate for properties of language for which the learner's hypothesis space is known to consist of a very small closed set of options: either heavily constrained by nature, as in the machinery for discovering the phonetic inventory of a natural language given the psychophysical properties of speech perception (18), or experimentally constrained, as in recent word-boundary learning experimentation, which vary the relative positioning of a small number of CV syllables (19). Similarly, the laboratory experiments we have reviewed, purporting to show that statistical-associative machinery can account for how humans extract concept-word mappings, limit the search space to a small set of fixed images that occur in temporal lockstep with an equally small set of nonsense sounds. However, actual word learning does not occur in anything resembling this picture-book setting but rather in uncontrolled, ever-changing environments in which each word occurs in different sentences with different exemplars inside different events, and only loosely intercalated with them in time and place. When the hypothesis space is very large, the role of memory becomes much more prominent.[†] Our findings cohere within a perspective in which memory limitations not only limit interpretive choices, but actually rescue the learning machinery from the pitfalls of false choice.

More specifically, our findings suggest that word learners are in one of three pertinent states during the acquisition of each word. In the initial state *A*, there is no mapping at all. Upon hearing the word in context, the machinery makes a (single) conjecture, thus passing into an interim state *B*. Now the machinery seeks confirmation, i.e., a new context at least weakly consistent with the one formed at *A*. If this step succeeds, the conjecture is further solidified as a confident hypothesis of the word's meaning. The crucial question is what happens when there is failure at *B*. What we see in the laboratory (Figs. 2 and 3) is that learners shift to a new conjecture, but this shift comes at some cost. Rather than returning to a state of semantic innocence (*A*), learners enter a memorial limbo (*B*), which leaves some residue of confusion that interferes with subsequent learning. This confusion is eliminated after a considerable delay, whereupon the machinery

[†]A trivializing alternative explanation is that our requirement that participants overtly state and record their interim conjectures reduced memory for prior conjectures. We investigated this issue separately in a manipulation in which participants responded overtly only during the second half of the experiment. This process had no impact on performance in the second half, compared with another set of participants who were overtly responding throughout.

returns to its initial state (A) and can again form a "first" conjecture (Exp. 3).

Conclusions

The proliferating memory burden inherent in storing multiple uncategorized contexts during the learning process suggests that recorded memories for past situations cannot be the central vehicle for constructing a lexicon. If it were, even more intractable problems arise with the evaluation procedure that the learner must instigate at some appropriate time (say, when 5 or 10 or 50 instances of "wode" or "blicket" have been heard and their associated situational contexts squirreled away "neutrally" in memory). What in particular could be true of all these wodish and blickettesque scenes? Prior findings from our laboratory have suggested that learners don't—as might be hoped—home in on a narrower and narrower interpretation as a consequence of such repeated encounters (e.g., "object" on the first encounter, "toy" on the second, "ball" by the third). Instead, learners do the reverse: moving from least to more inclusive—that is, more abstract—interpretations as trials and, therefore, differences in observations increase. That is, one can always find "some" similarity among a set of disparate situations if one's conjecture becomes vague and general enough. In any set of human encounters, someone is almost always "looking" and there is just about always some "thing" in view. If the idea is to parse out of scenes that which is common to all of them, this increasing abstractness (cross-situational generalization of the bad kind) is bound to happen. And this is exactly what participants do in experimental conditions that encourage remembrance of things past, namely massed trials in which exposures to new words are not interspersed with other words, new or known (14). Yet everything we know about vocabulary growth tells us that word learning does not work this way; rather, the early vocabulary is highly concrete, even where learner age varies (12). One-trial learning presupposes strong restrictions on the hypotheses that will even be entertained (e.g., ref. 18) and about conditions on the interpretation of what is being observed. Such conditions of "relevance" or "newsworthiness" apply everywhere in human social intercourse, not only in the special situation of word learning.

Such an analysis of the problem would suggest that word learners would be well advised not to consult specific memories of past situations during word learning. Rather, a process in which a single conjecture is stored, defeasible only if countermanded by a failure during the confirmation phase, is a more reasonable approach, and one that is strongly supported by the experimental findings reported here. Learners interrogate scenes and analyze them for a plausible meaning, forming a single conjecture that, unless explicitly and rapidly countermanded, can last a lifetime. So conceived, learners acquire the meanings of concrete terms "from" observation, but only when the shoe fits.

Methods

Participants. Undergraduates from the University of Pennsylvania and Harvard University participated for course credit or $10/h: 37 participants in Exp. 1, 64 in Exp. 2, and 35 in Exp. 3.

Stimuli. Stimuli for Exp. 1 were 288 40-s muted videos (vignettes) of parent–toddler (ages 12–15 mo) interactions in natural settings, of the 48 most frequent nouns and verbs (24 each) in the corpus. Vignettes were aligned so that 30 s into the video, the parent uttered the target word (at which a beep was heard). Previous studies found this duration to be sufficient to understand the gist at the moment the target word was uttered (13, 14). Exps. 2 and 3 target stimuli were a subset of these vignettes (40 total) and were examples of a parent uttering one of eight common nouns (five vignettes each of *bag*, *ball*, *book*, *horse*, *necklace*, *nose*, *phone*, and *shoe*). A spoken nonsense word (e.g., "mipen") was spliced into the silent audio instead of the beep, with the same timing. Each nonsense word corresponded to a target word (e.g., *mipen = bag; vash = shoe*), thereby permitting cross-situational word learning. *SI Text* describes selection criteria for the 40 target vignettes, which were based on the results of Exp. 1.

Design of Exps. 2 and 3. For the five vignettes of each target word, occurrence of the HI vignette was manipulated relative to the four LI vignettes: it occurred First, Middle, Last, or was Absent. The four LI vignettes were placed in a fixed random order relative to the position of the HI vignette. In the HI Absent condition, the four LI vignettes were followed by a repetition of the first LI vignette.

Presentation of target words was intermixed (distributed presentation) such that each participant saw the first instance of each target word one after the other, followed by all of the second instances, and so forth. In Exp. 2, participants were randomly assigned to one of the four conditions (HI First, Middle, Last, or Absent) and one of two stimulus orders (forward or reverse). Exp. 3 was the same except that only HI Middle and HI Last lists were run.

Procedure of Exps. 2 and 3. See *SI Text*, including procedure for Exp. 1, which was similar to Exps. 2 and 3. Participants were tested in small groups in a room with a video projector. Participants were told they would watch videos of parents interacting with their children, with parents uttering 1 of 12 common words. The sound would be absent but the "mystery word" would be indicated by a corresponding spoken nonsense word, played at the exact moment that the parent uttered the mystery word. Participants had to "figure out which English words correspond to these twelve nonsense words, given the scenes that you watch in these videos." After each vignette, participants wrote their guess and rated confidence from 1 (low) to 5 (high). Participants could not review their previous responses. At the end, participants were given a list of the nonsense words and recorded a final conjecture and rating for each word.

Exp. 3 was identical to Exp. 2, except subjects participated in two sessions separated by 1 to 3 d. During the first session, participants viewed the vignettes leading up to but not including the HI vignettes. In the second session they viewed the HI vignettes and any remaining LI vignettes.

ACKNOWLEDGMENTS. This work was funded by National Institutes of Health Grant 1-R01-HD-37507.

1. Carey S (1978) *Linguistic Theory and Psychological Reality*, eds Halle M, Bresnan J, Miller GA (MIT Press, Cambridge), pp 264–293.
2. Baldwin DA (1993) Infants' ability to consult the speaker for clues to word reference. *J Child Lang* 20:395–418.
3. Gleitman LR, Cassidy K, Nappa R, Papafragou A, Trueswell JC (2005) Hard words. *Lang Learn Dev* 1:23–64.
4. Dollaghan CA (1985) Child meets word: "Fast mapping" in preschool children. *Child Dev* 58:1021–1034.
5. Heibeck TH, Markman EM (1987) Word learning in children: An examination of fast mapping. *Child Dev* 58:1021–1034.
6. Yu C, Smith LB (2007) Rapid word learning under uncertainty via cross-situational statistics. *Psychol Sci* 18:414–420.
7. Smith LB, Yu C (2008) Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition* 106:1558–1568.
8. Xu F, Tenenbaum JB (2007) Sensitivity to sampling in Bayesian word learning. *Dev Sci* 10:288–297.
9. Vouloumanos A (2008) Fine-grained sensitivity to statistical information in adult word learning. *Cognition* 107:729–742.
10. Smith K, Smith ADM, Blythe RA (2010) Cross-situational learning: An experimental study of word learning mechanisms. *Cogn Sci* 35:480–498.
11. Gentner D, Boroditsky L (2001) *Language Acquisition and Conceptual Development*, eds Bowerman M, Levinson S (Cambridge University Press, Cambridge), pp 215–256.
12. Snedeker J, Geren J, Shafto CL (2007) Starting over: International adoption as a natural experiment in language development. *Psychol Sci* 18:79–87.
13. Snedeker J, Gleitman LR (2004) *Weaving a Lexicon*, eds Hall DG, Waxman SR (MIT Press, Cambridge), pp 257–293.
14. Gillette J, Gleitman H, Gleitman L, Lederer A (1999) Human simulations of vocabulary learning. *Cognition* 73:135–176.
15. Papafragou A, Cassidy K, Gleitman L (2007) When we think about thinking: The acquisition of belief verbs. *Cognition* 105:125–165.
16. Csibra G (2010) Recognizing communicative intentions in infancy. *Mind Lang* 25: 141–168.
17. Booth AE, Waxman SR (2002) Word learning is "smart": Evidence that conceptual information affects preschoolers' extension of novel words. *Cognition* 84:B11–B22.
18. Liberman AM, Cooper FS, Shankweiler DP, Studdert-Kennedy M (1967) Perception of the speech code. *Psychol Rev* 74:431–461.
19. Saffran JR, Aslin RN, Newport EL (1996) Statistical learning by 8-month-old infants. *Science* 274:1926–1928.
20. Steiger JH, Shapiro A, Browne MW (1985) On the multivariate asymptotic distribution of sequential chi-square statistics. *Psychometrika* 50:253–264.