Cognitive Psychology 66 (2013) 126–156

Contents lists available at SciVerse ScienceDirect

Cognitive Psychology

journal homepage: www.elsevier.com/locate/cogpsych

Propose but verify: Fast mapping meets cross-situational word learning

John C. Trueswell *, Tamara Nicol Medina, Alon Hafri, Lila R. Gleitman

Department of Psychology, University of Pennsylvania, 3720 Walnut St., Philadelphia, PA 19104, USA

ARTICLE INFO

Article history:
Accepted 5 October 2012
Available online 9 November 2012

Keywords:
Cross-situational word learning
Fast-mapping
Statistical learning
One-trial learning

ABSTRACT

We report three eyetracking experiments that examine the learning procedure used by adults as they pair novel words and visually presented referents over a sequence of referentially ambiguous trials. Successful learning under such conditions has been argued to be the product of a learning procedure in which participants provisionally pair each novel word with several possible referents and use a statistical-associative learning mechanism to gradually converge on a single mapping across learning instances [e.g., Yu, C., & Smith, L. B. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science, 18*(5), 414–420]. We argue here that successful learning in this setting is instead the product of a one-trial procedure in which a single hypothesized word-referent pairing is retained across learning instances, abandoned only if the subsequent instance fails to confirm the pairing – more a 'fast mapping' procedure than a gradual statistical one. We provide experimental evidence for this *propose-but-verify* learning procedure via three experiments in which adult participants attempted to learn the meanings of nonce words cross-situationally under varying degrees of referential uncertainty. The findings, using both explicit (referent selection) and implicit (eye movement) measures, show that even in these artificial learning contexts, which are far simpler than those encountered by a language learner in a natural environment, participants do not retain multiple meaning hypotheses across learning instances. As we discuss, these findings challenge 'gradualist' accounts of word learning and are consistent with the known rapid course of vocabulary learning in a first language.

© 2012 Elsevier Inc. All rights reserved.

* Corresponding author. Address: Department of Psychology, University of Pennsylvania, 3720 Walnut Street, Solomon Lab Bldg., Philadelphia, PA 19104, United States. Fax: +1 215 573 9247.
E-mail address: trueswel@psych.upenn.edu (J.C. Trueswell).

0010-0285/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.cogpsych.2012.10.001

1. Introduction

Common sense tells us that when children learn the meanings of some initial set of words in their language, they must be relying heavily on their observation of the immediately co-occurring context; it must be that the presence of cats when /kæt/ is uttered plays a dominant causal role in establishing that /kæt/ means 'cat' in English. Similarly, adults who find themselves immersed in an unfamiliar language community must at least initially rely on similar observational evidence to break into the language. Despite the inevitability of this idea about the conditions for early word learning, it has been remarkably difficult to specify how such a procedure of contextualized observation might work in practice. Arguably the biggest hurdle is that words are necessarily uttered in complex situational environments, and thus are subject to a wide variety of alternative plausible interpretations. How, then, could learners determine what the interlocutor is referring to? Even if the intended referent for a word can be effectively recovered, the speaker's intended conceptual characterization of that referent often remains ambiguous, yet it is this characterization that should determine a word's meaning (Chomsky, 1957; Fodor, 1983; Gleitman, 1990; Quine, 1960; *inter alia*).

Experimental research on this topic has shown that learners possess conceptual and referential biases that redress some of these problems, allowing learners significant narrowing of the hypothesis space of plausible word-meaning mappings. For instance, both toddlers and adults are more likely to assume that a new word refers to a whole object in view than to its parts (Markman, 1989) and more likely to categorize it by its shape than its size or color (Landau, L. Smith, & Jones, 1988). Moreover, research over the past 40 years documents that, after an initial "true novice" stage, observational cues do not operate within a stand-alone procedure, but are supplemented and coordinated with multiple linguistic and social cues (e.g., Gleitman, Cassidy, Nappa, Papafragou, & Trueswell, 2005). These include phonetic (e.g., Gervain, Nespor, Mazuka, Horie, & Mehler, 2008), syntactic (Arunachalam & Waxman, 2010; Fisher, Hall, Rakowitz, & Gleitman, 1994; Landau & Gleitman, 1985; Naigles, 1990; Snedeker & Gleitman, 2004; Trueswell & Gleitman, 2007; Yuan & Fisher, 2009), distributional (Carey, 1978; Maratsos & Chalkley, 1981; Newport & Aslin, 2004; Saffran, Aslin, & Newport, 1996) and social-attentive properties (e.g., Baldwin, 1991; Baldwin, 1993; Gillette, Gleitman, Gleitman, & Lederer, 1999; Jaswal, 2010; Nappa, Wessell, McEldoon, Gleitman, & Trueswell, 2009; Papafragou, Cassidy, & Gleitman, 2007) of the situations in which conversation takes place.

Nevertheless, every known theory of vocabulary growth posits an initial novice stage during which observation of the passing scene is the primary, even sole, source of information about a new word's meaning. The priority of observation as a learning cue follows from the fact that a true novice cannot – without prior learning based on observation – exploit syntactic or distributional cues that may be present in the input situation. Without some knowledge of the specifics of the exposure language to provide a zoom lens highlighting the perspective the speaker is adopting, the situational context often remains surprisingly indeterminate. Rampant indeterminacy during this initial novice stage has been illustrated in the laboratory in both adults and young children by degrading the participants' input such that observation of context is the only available cue (e.g., Gillette et al., 1999; Medina, Snedeker, Trueswell, & Gleitman, 2011; Piccin & Waxman, 2007; Snedeker & Gleitman, 2004). In this procedure (known as the Human Simulation Paradigm, HSP), participants watch videotaped segments (approximately one minute in duration) of parents in everyday contexts speaking to their 15–18 month old offspring, using the most common words in a sample of maternal corpora. The sound is turned off in the video, and only a beep or a nonsense word indicates the moment at which a particular word had been uttered. Participants' task is to guess these "mystery words" from this social and visual information alone, effectively placing them in the position of true novices in this initial stage of word learning. Participants in these tasks fail rather strikingly at determining the words that parents uttered to their toddlers in these videotaped natural interactions. For example, Medina et al. (2011) found that participants achieve accuracy scores of over 50% for only 7% of a set of randomly chosen items and instances, even though all test words were among the most frequently occurring nouns and verbs in maternal speech to infants. Children performing this same HSP task show strikingly similar difficulty in identifying intended referents, even though they might be expected to be better versed in identifying parental cues to reference (Medina et al., 2011; Piccin & Waxman, 2007).

Given the documented difficulty of learning word-referent pairings from a single observation, how can the indeterminacy problem be overcome? The solution most frequently offered by researchers is that a learning mechanism that compares multiple learning instances for a given word can mitigate or even erase the indeterminacy of any single instance. Indeed, there have been numerous experimental demonstrations that, even with referentially ambiguous learning instances, toddlers, children and adults are quite good at converging on the intended referent for a word across learning instances (e.g., K. Smith, A. Smith, & Blythe, 2011; Yu & L. Smith, 2007, 2011). Here we examine in detail the mechanism that supports this learning. As we discuss below, the common assumption is that learners achieve cross-situational learning by keeping track of multiple hypotheses about a word's meaning across successive learning instances, and gradually converge on the correct meaning via an intersective statistical process. Yet, contrary to this assumption, we report below that when learners are placed in the initial novice state of identifying word-to-referent mappings across learning instances, evidence for such a multiple hypothesis tracking procedure is strikingly absent. "Cross-situational" learning appears to be less cross-situational than has usually been conjectured.

1.1. Cross-situational comparison

Plausibly, listeners who appreciate the indeterminacy of observation will not jump to a conclusion about word meaning on first observation, but rather will hold the choice in abeyance until evidence from further exposures has accumulated. On most accounts, learners pursue such a strategy by storing the multiple conjectures generated on each observation of a word in its situational context, and then comparing among these conjecture-situation lists. Over time and exposures, there will presumably be increasing co-occurrence between a word and its target meaning and decreasing co-occurrence between the word and other potential meanings (e.g., Frank, Goodman, & Tenenbaum, 2009; Osgood, Suci, & Tannenbaum, 1957; Siskind, 1996; Xu & Tenenbaum, 2007; Yu & L. Smith, 2007; Yu, 2008). Indeed, extant video corpora of parent–child discourse (e.g., Gillette et al., 1999; Medina et al., 2011) seem to make such a procedure available. For example, in the Medina et al. (2011) corpus, a book is present in all of several randomly selected video clips in which the parent uttered "book", while bowls and dogs and spoons, etc., were visible in only one or two of them. The key to all current models of cross-situational word learning is that they capitalize on such repeated co-occurrences of a word and its referent by formulating an associative learning process that tracks the frequencies of potential referents across all of the contexts in which the word is used.

The recent literature tries to operationalize this cross-situational learning procedure and test it in controlled experimental settings. To achieve some experimental control over the saliency and size of the conjecture set for a word, most studies use a set of artificial stimuli and a fixed number of exposures, but vary the degree of indeterminacy within and across trials. For example, in Yu and L. Smith (2007), participants heard a list of nonsense words (2–4 words) and simultaneously viewed a small set of photographic images (2–4 pictures of objects) on a computer screen. Participants knew that the spoken nonsense words labeled each object on the screen (e.g., with four objects one might hear "mipen, dax, moop, blang"), but otherwise were not informed about the experimenter's intended pairings of words and referents. Here referential ambiguity exists on each trial, and the only way of solving the mapping problem is to retain and compare information across trials.

The central finding in this and related experiments was that participants' mapping success was a function of the co-occurrence statistics for each word. Specifically, after the learning trials were complete, participants were tested on their understanding of the nonsense words using a 4-alternative forced choice paradigm in which they heard a nonsense word paired with the target referent and three distracters. Group performance was well above chance and a function of the referential ambiguity present during learning: items learned under higher degrees of referential ambiguity showed lower, albeit above chance, performance as compared to items learned under conditions of lower referential ambiguity. Subsequent studies have found similar high rates of successful cross-situational learning under conditions of referential ambiguity (e.g., Ichinco, Frank, & Saxe, 2009; Kachergis, Yu, & Shiffrin, 2010; Klein & Yu, 2009).

Yu and L. Smith (2007) concluded that although a range of learning procedures could explain their results, any such model would need to include the simultaneous consideration of multiple

word-to-meaning hypotheses, either through (a) a large association network connecting forms to meanings; (b) a more complex network that includes inhibition among competing associations; or (c) statistical learning that explicitly compares alternative hypotheses (p. 419). Yu (2008) developed an associative computational model of early word learning that simulates the results of Yu and L. Smith (2007) in which multiple referential alternatives were tracked for each word across learning instances. Relatedly, Frank, Tenenbaum, and colleagues have developed Bayesian models of cross-situational word learning that simultaneously consider all possible word-meaning mappings to capture other cross-situational word-learning findings (e.g., Frank, Tenenbaum, & Fernald, in press; Frank et al., 2009).

1.2. Propose but verify: Single conjectures from multiple exposures

There is an alternative way that the learner could capitalize on repeated co-occurrences of a word and its referent without having to keep track of co-occurrence frequencies of multiple potential referents from each of the contexts in which the word occurs. The learner could make a single conjecture upon hearing the word used in context and carry that conjecture forward to be evaluated for consistency with the next observed context. If the guess is "confirmed" by consistency with the succeeding observation, the learner will further solidify the word meaning in memory. If the guess is inconsistent with the succeeding observation, the learning machinery will abandon this interpretation and postulate a new one – which can be carried forward, in its turn, for subsequent confirmation or rejection. It follows from this account that the more consistent the co-occurrence statistics are in the input, the more likely the learner is to make a correct conjecture at some point and then to confirm it. (If some malign parent alternates by uttering "rhinoceros" on one exposure and "elephant" on the next, learning will not occur.) Notice that this hypothetical learner, unlike the hypothetical associative learner, need not actually track the cross-trial statistics in order to build the correct mappings between words and referents. We will call this learning strategy ***propose-but-verify***, reflecting the need for the learner to maintain a single hypothesized meaning that is then tested at the next learning instance for that word.

Initial support for the propose-but-verify procedure comes from Medina et al. (2011). Here participants attempted to learn words by watching muted video clips of parent–child interactions (i.e., the HSP procedure of Gillette et al., 1999). This visual-contextual situation was far more complex and variable, therefore, than in the cross-situational experiments discussed earlier, which had used a set of photographic images, unvarying across trials, as the sole example of each word's context. With these more naturalistic stimuli, Medina et al. found that when participants guessed a word (identified as a nonsense word such as "mipen" or "pilk") correctly in the first context in which they observed it, they tended to maintain this interpretation in later contexts in which the word occurred. In contrast, when participants guessed wrongly in the initial context, they rarely recovered across later learning instances. Rather, they made new but still incorrect guesses for each of the subsequent learning instances for that word. Crucial evidence that learners were tracking just a single word-meaning hypothesis across learning instances came from more detailed contingent-response analyses related to this phenomenon. Specifically, when learners had guessed incorrectly on any given learning instance, mean performance on the very next learning instance was extremely poor (11% correct), with an accuracy almost identical to the performance of other participants who had viewed the same videos in isolation and hence had no benefit of cross-situational comparison (9% correct). This suggests that learners had no memory of the plausible alternative referents that arose on the previous learning instance. Strikingly, this held even when the previous learning instance had been highly informative to most other learners in the experiment. For example, for a learning instance on which the majority (65%) of participants had guessed the correct referent, the remaining minority (35%) who had guessed incorrectly then went on to perform quite poorly on the next trial, just as poorly as those who had no benefit of a previous learning instance. If participants had been tracking multiple referential alternatives (as e.g., Yu, 2008, and Yu & L. Smith, 2007, predict), it seems plausible that many of these incorrect participants would have stored the correct referential alternative that most other participants had guessed on that same trial. But instead, no memory of this plausible alternative was observed in their responses on the next learning instance – even though this learning instance provided evidence that also supported this meaning.

1.3. Comparing the accounts

We have just sketched two accounts of how learners might solve the indeterminacy problem cross-situationally in early word learning. In the first, learners would store multiple conjectures and, in the presence of accumulated data, implicitly compare among them via a gradual associative procedure (Yu & L. Smith, 2007). Medina et al. (2011) suggested that such a storage-comparison process could not work owing to the size and scope of the conjecture sets that real life would actually make available. Learners were instead proposed to use a procedure that remembers only its current conjecture, forgetting the observational settings that engendered it.

The experimental findings that have been provided as support for these alternative accounts come from studies using strikingly different test stimuli. As is so often the case, the experimental strategies have gone in two directions: One series of studies opts for control of the size and saliency of items in the stimulus set, but at cost of unknown oversimplification of the natural case (e.g., Ichinco et al., 2009; Kachergis et al., 2010; Klein & Yu, 2009; Yu & L. Smith, 2007, 2011); the other series tests learning in the buzzing, blooming confusion of real parent–child interaction, but at cost of leaving unexplicated the nature of visual-contextual environment leading to successful word learning (Medina et al., 2011). Moreover, there is an observational–motivational difference between these two laboratory situations whose influence is hard to assess; namely, in the HSP paradigm used by Medina et al. we cannot be sure that what is salient for a participant-observer of a video is just as salient as for the child being filmed in the video (for discussion of this objection, see L. Smith, Yu, & Pereira, 2009): perhaps the participants in Medina et al., who were placed in the position of a third-person observer rather than the second-person addressee of the utterance, did not actually notice the target referent in one or more of the video clips they watched, or did not consider it salient enough to maintain in memory as a potential referent.

Finally, and most importantly, the word learning experiments involving artificial stimuli such as Yu and L. Smith (2007) all purport to show that multiple meaning hypotheses are stored across exposures to a word, comparing among them along the way, and then settling on one of them at the end. Yet none of these experiments actually examined how learning accuracy unfolded across learning instances. Rather, only final performance was evaluated. As suggested by Medina et al. (2011), one must examine the sequence of responses across learning instances to differentiate such word-learning accounts from simpler ones. For instance, participants might form a single hypothesis upon some early exposure, and then use the information in subsequent exposures only to confirm or disconfirm this conjecture. Such a strategy could account for the same outcomes in artificial stimuli studies that are usually described as a multiple-conjecture comparison process. Under this latter perspective, "cross situational learning" has become, in effect, one-trial learning with a confirmation check, as in the propose-but-verify proposal of Medina et al.

To the best of our knowledge, only one published research project other than Medina et al. (2011) examined learning patterns across multiple exposures to a new word. K. Smith et al. (2011) used artificial learning environments like those used by Yu, L. Smith, and colleagues (e.g., Yu & L. Smith, 2007) and recorded word meaning accuracy after each learning instance. However, K. Smith et al. never tested whether a propose-but-verify learning procedure would fit their data (nor offered data inconsistent with this procedure[1]) and instead only compared models that stored multiple meaning conjectures for each word. K. Smith et al. allude to a propose-but-verify model but never tested it, suggesting that such a learning procedure would only be used as a last resort, under conditions of extreme referential uncertainty; under all simpler conditions, learners would be expected to track multiple conjectures.

In sum, all but one of the cross-situational experiments that have used simplified artificial stimuli have not examined the underlying learning procedure that gave rise to successful performance at study end; the experiment that did (K. Smith et al., 2011) did not test if a single-hypothesis (propose-but-verify) procedure would support the patterns of learning observed across trials. It is possible

[1] We refer here to the results involving distributed learning, in which learning instances of different words were intermixed rather than artificially grouped together by word (blocked learning). Under the more natural conditions of distributed learning, K. Smith et al. (2011) either were unable to fit any multiple-conjecture learning model to their data, or found that a very simple multiple-conjecture model was a better fit than a model that performed no learning at all. A propose-but-verify (single-conjecture) model was never examined to see if it offered a better fit.

that under the simplified conditions of artificial learning stimuli, participants track multiple hypothesized meanings for each word, or it could just as well be that single-hypothesis learning tracking is the norm, observed even under simplified conditions.

1.4. The present experiments

We present here three experiments that explicitly tested whether learners take into account a set of potential word meaning hypotheses from one learning instance to the next – to be used for cross-situational comparison – or whether they make a single conjecture upon hearing the word used in context and carry only that conjecture forward to be either confirmed or disconfirmed in terms of consistency with the new context (propose-but-verify). The tests follow a straightforward logic, used in Medina et al. (2011) but applied here to the artificially controlled test items, and item-to-nonse-word pairings, that have been employed in the literature on this topic. In particular, consider a sequence of two learning trials illustrated in Fig. 1 below, in which the correct meaning for "zud" happens to be *bear*. If on the first learning instance, a participant has incorrectly selected *door* as the meaning of "zud", then according to the propose-but-verify strategy, the participant should not store the alternative word meanings from that instance (*hand, dog, ball*, and [the correct referent] *bear*). Thus when encountering the next instance of "zud" (on the right of Fig. 1), the participant should select randomly among the referents, *even though one of the alternatives is a bear,* which appeared in the previous learning instance. If on the other hand alternative hypotheses are being tracked, the participant should be above chance at selecting the bear on this second learning instance. He or she would have memory of past alternatives, even when not chosen. It is possible that such a simple cross-situational learning strategy should occur only under very high levels of referential ambiguity – under simpler referential contexts perhaps people can track multiple word-to-meaning pairings (cf. K. Smith et al., 2011). To test this, Experiments 2 and 3 reduced referential ambiguity to the lowest extent possible – namely, two alternatives – while still keeping the learning trials ambiguous.

A finding that learners recover from previously incorrect guesses at rates above chance would suggest that they remembered the target referent from the previous learning instance even if they had not indicated that referent as their response. However, a finding that learners remain at chance following a previously incorrect guess would demonstrate that they did not recognize the recurrence of the target referent across contexts, as expected by the propose-but-verify learning procedure.

2. Experiment 1

2.1. Introduction

Here we examine whether a series of learning instances containing a high degree of referential uncertainty (5-referent alternatives, Fig. 1) generates a pattern of gradual learning, at least in the

Fig. 1. Example sequence of two learning trials for the word "zud", which meant 'bear'. In both instances, five alternatives are displayed. Note that different examples of bears appear, that only one entity is labeled on each trial (i.e., only one nonsense word is heard), and that other learning trials for other words intervene.

aggregate. We will then use contingent-response analyses, described in Section 1.2 above, to determine whether learning, if observed, arose from a process of retaining and testing multiple hypotheses across learning instances, or from a process in which a single hypothesis is retained and tested across learning instances.

2.2. Method

2.2.1. Participants

Fifteen undergraduates from the University of Pennsylvania participated for course credit or $10/h. All were native speakers of English. All provided written informed consent before participating. Participation lasted approximately 20–30 min, including time for directions.

2.2.2. Stimuli

In order to create visual stimuli for this experiment, five different but readily identifiable photographic images were selected from clipart.com or photographed anew as example referents for each of twelve common object labels (*ball, bear, bird, book, car, cat, dog, door, eye, fish, hand,* and *shoe*), resulting in 60 images in total. Any background material behind the object of interest was removed so that all objects appeared on a white background.

We then paired each category with a single nonsense word, resulting in twelve pairings (e.g., *ball* = "smick"; *bear* = "zud"). The words followed the phonological rules of English and were intended not to be similar to common English object labels.

The visual display for each trial consisted of five image referents arranged symmetrically around a crosshair (see Fig. 1). Henceforth learning instances with five referential alternatives are designated "Low Informative" (LI). Each of the five images was selected from a different category. The visual display was accompanied by a pre-recorded object-labeling utterance by a female speaker that referred to one of the five objects, of the form "Oh look, a _____!", ending with a unique nonsense word (e.g., "smick").

2.2.3. Apparatus

A Tobii 1750 remote eyetracking system was used for stimulus presentation and data collection. This system performs binocular tracking using optics embedded in a flat panel monitor with a display size of 33.5 (width) × 26.75 (height) cm (31.2 × 26.9 deg visual angle at a viewing distance of 60 cm). Two laptop computers running the Windows XP operating system were used to control the system: one displayed stimuli on the eyetracker screen (via E-Prime experiment software) and the other collected eyegaze data from the eyetracker (via the TET-server software developed by Tobii Technology). Both laptops were disconnected from the internet to increase timing accuracy. The data sampling rate was a consistent 50 Hz, and the spatial resolution of the eyetracker is approximately 0.5–1.0 deg visual angle, including corrections for head drift and small head movements. At a 60-cm viewing distance, the Tobii 1750 has a tolerance to head motion of about 30 × 16 × 20 cm. The system recovers from a complete eyetracking failure in <100 ms.

Each of the five referent images subtended approximately 6.1 × 6.4° visual angle. Each was located between 4.5° and 5.4° from the central crosshair and between 1.6° and 3.4° from any neighboring referent image. This allowed for accurate discrimination of gazes to each referent image.

2.2.4. Procedure

Participants were tested individually. Each was seated approximately 60 cm from the Tobii screen, and the experimenter adjusted the angle of the screen as necessary to obtain robust views of both eyes, centered in the tracker's field of view. The Tobii ClearView default 5-point calibration scheme was then used to obtain an accurate track of both eyes. If the calibration data did not meet the default criteria of the ClearView software or if it was incomplete (fewer than 5 points), the calibration was repeated. Participants typically required only one calibration.

The participant was then given instructions for the experiment. They were told that they would be seeing a series of trials on which they would see one or more pictures of objects on the screen and would hear a single "mystery" word. They were informed that there would be 60 trials in total, and

that over the course of the experiment they would hear 12 different mystery words, each with a different meaning. They were to figure out what each mystery word meant. They were instructed to "click on one of the objects that you think the word might be referring to". No feedback was given to participants about the correctness of their answers. At the end of the experiment, participants were presented with the audio of each of the nonsense words and asked to say aloud what they thought the word meant. The experimenter then wrote down each spoken response. Along with their eye movements, participants' clicking responses, including their timing and accuracy, were logged by the computer. Spoken responses from the end of the experiment are not reported below because chance performance cannot be determined from such data. Needless to say, the vast majority of responses (over 97% in Experiments 1 and 2 and over 95% in Experiment 3) were of basic level categories such as "cat" and "door", and not superordinate ("animal" or "pet") or subordinate categories ("Manx" or "tabby.").

2.2.5. Experimental design

The twelve nonsense words (referring to the twelve object categories) occurred in five trials each, resulting in a total of 60 trials. The co-occurrence frequency of a word and a target referent (e.g., a picture of a cat) was 100% – five out of five trials. On each of these five trials, the particular picture from the referent category was different (e.g., a white cat, a Siamese cat).

Each of the five pictures was randomly assigned to appear in an additional four trials as a distractor referent, for a total of 20 additional appearances of each object category over the course of the experiment. Importantly, however, each object category was restricted from appearing more than twice with a particular nonsense word as a distractor (e.g., if "smick" meant ball, then across the five "smick" trials, there was always a ball, but no more than two of those trials contained a book). Thus, the co-occurrence frequency of a word with a distractor referent was maximally 40% – two out of five trials.

Presentation of the words was intermixed (i.e., a distributed presentation) such that each participant saw the first learning instance for each word one after the other, followed by all of the second instances, etc. Within each round of learning instances (e.g., the group of first learning instances), the order of the words was always the same pseudo-random order. A second presentation list was created in which order of trials within each block was approximately reversed, such that for those words which were selected as target words in the later experiments (eight of 12 words), the same Target–Target–Filler presentation was preserved (see Section 3.2.3 of Experiment 2 below for further clarification). Participants were randomly assigned to one of the two experimental lists.

2.3. Results and discussion

2.3.1. Learning curve

Fig. 2A plots the average proportion of correct clicking responses over the five learning instances, collapsing across each word. The mean results suggest that learning was difficult but not impossible. Indeed, a growth curve analysis, using a multi-level logit model of accuracy data, showed a reliable increase in accuracy across instances (see Table 1).

2.3.2. Accuracy-contingent clicking responses

Given that reliable learning patterns are observed in our data, we now ask what kind of learning underlies this pattern. As discussed in the introduction, traditional statistical word learning models posit that participants track all word-to-referent hypotheses across learning instances. If true, participants should show memory of the alternative correct referent even when that referent had not been selected (clicked on) previously; that is, they should show above chance performance at selecting the correct alternative on the next learning instance since it too has the correct referent present in the display. We tested this hypothesis by calculating the average proportion of correct responses on instances 2–5, split by whether the participant had been correct or incorrect on the previous learning instance for that word (Fig. 2B). That is, for learning instance N, we graph average accuracy as a function of correctness on instance $N-1$. We have collapsed across instances 2–5 rather than plotting each separately because of the relatively low number of observations that would result from such a

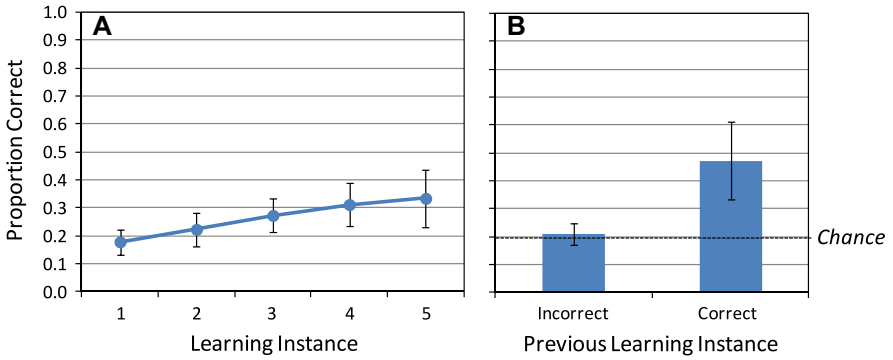

Fig. 2. Mean proportion of correct responses. Participant means. Error bars indicate ± 95% C.I. (A) As a function of learning instance. (B) As a function of whether the participant had been correct or incorrect on the previous learning instance for that word. Number of participants contributing to bars from left to right: 15 out 15, 15 out of 15 (Experiment 1).

Table 1
Learning in the aggregate: Effect of learning instance on accuracy (Experiment 1).

Effect	Estimate	S.E.	*z*-value	*p*-value
Intercept	−1.16	0.21	−5.51	<.0001*
Instance	0.21	0.07	3.25	.001*

Note: Results of multi-level logit model with crossed random slopes and intercepts for both Participants and Items. The lmer function in *R* was used [syntax: Accuracy ~ 1 + Instance + (1 + Instance|Participant) + (1 + Instance|Item)]. Model was a better fit than a corresponding model that did not include Instance as a fixed effect, based on a chi-square test of the change in −2 restricted log likelihood ($\chi^2(1) = 7.55$, $p = .006$). Instance variable was centered.
* Statistically significant.

division of the data (though see Experiment 3 below for such an analysis on a larger number of subjects).

This figure plots the average of participant means, with error bars indicating a ± 95% confidence interval. Thus, if the error bar does not touch the .20 proportion line (which in this case is 1-in-5 chance performance), then participants were found to behave above chance, as tested by a one sample *t*-test (2-tailed) on participant means. The same patterns of significance were also found in tests of item means against chance levels. We therefore discuss significance based on the error bar patterns and do not report these other tests of chance performance.[2]

As can be seen in the figure, participants were above chance only after guessing correctly for a given word. After guessing incorrectly, participants randomly selected a referent, resulting in 1 out of 5 (.20) performance. Thus, even though the Target referent (e.g., a *bear*) had been present the last time participants heard the word in question (i.e., "zud") and it was present again on the current instance, participants showed no sign of remembering this fact if they had not selected the bear previously. Such a pattern is expected if participants remember only the current hypothesized meaning for a word – discarding elements of the situation that engendered their hypothesis – and are seeking to verify their single hypothesis by its relevance in the current learning instance. Learning is not perfect, however, even when the subject guesses correctly on one trial and confirms it on the next (less than 100% correct performance in this ideal situation). This suggests that participants sometimes fail to recall even their own hypothesized meaning.

[2] Unless otherwise noted, subject and item mean tests confirmed the error bars in all accuracy contingent analyses reported below in Experiments 2 and 3. Also, given the nature of contingency analyses, it is possible for a subject not to contribute to a subject mean. We therefore include in the figure caption the number of subjects out of the total that contributed to the mean in all accuracy-contingent response figures in this paper.

2.3.3. Eye movements

It is possible that although participants' clicking behavior indicated that they used only the current hypothesized meaning to inform responses, their eye movements would reveal some implicit memory for the alternate hypotheses. If this is the case, we would expect looks to the Target referent to exceed looks to a Competitor even when the participant had guessed incorrectly on the previous learning instance. We examined this possibility by comparing the average proportion of looks to the Target and a randomly selected Competitor referent in response to hearing the target word.[3] We compared learning instances on which the participant had responded correctly on the previous learning instance to learning instances on which the participant had responded incorrectly on the previous learning instance (Fig. 3). Fig. 3A plots the Target and Competitor looks separately, whereas Fig. 3B plots the difference, i.e., the Target Advantage Score (TAS), reflecting the proportion of time spent looking at the Target minus the proportion of time spent looking at the randomly selected Competitor. A positive TAS indicates a preference to look at the Target whereas a negative TAS indicates preference to look at the Competitor.

As can be seen in the figure, Target looks exceeded Competitor looks only when the participant had been correct on the previous instance. When the participant had been incorrect on the previous learning instance, the average proportion of Target looks and Competitor looks were very similar. This suggests that participants had no implicit memory that the previous learning instance included the Target referent. Instead participants recalled only the current working hypothesis about the word's meaning.

These conclusions found support in multi-level linear modeling of the eye movement data. For each trial, Target Advantage Scores (TASs) were calculated within four different time windows: (1) 0–499 ms; (2) 500–999 ms; (3) 1000–1499 ms, and (4) 1500–1999 ms from word onset.[4] For each time window, this trial-level data was entered into two different multi-level models, both having crossed random intercepts for Subjects and for Items (Baayen, Davidson, & Bates, 2008). The first model had no fixed effects (i.e., the 'null model'). The second model was the same except that a single fixed effect was added: Previous-Instance Accuracy (Correct vs. Incorrect). Significance of Previous-Instance Accuracy was assessed based on whether this second model reliably improved the fit of the data as compared to the null model based on a chi-square test of the change in −2 restricted log likelihood (Steiger, Shapiro, & Browne, 1985). Using this method, it was found that Previous-Instance Accuracy was reliable in the second ($\chi^2(1) = 10.2$, $p = .001$) and third ($\chi^2(1) = 21.9$, $p < .001$) time windows, and marginally significant in the fourth time window ($\chi^2(1) = 3.89$, $p = .05$). Moreover, in the third time window the TAS was reliably positive when the participant had been correct on the previous trial (est. mean = 1.16, $SE = 0.43$, $t = 2.70$, $p = .02$) but not when the participant had been incorrect (est. mean = −0.13, $SE = 0.14$, $t = -1.82$, $p = .41$).[5]

3. Experiment 2

3.1. Introduction

The results thus far are consistent with the propose-but-verify model of word learning, rather than the models that track multiple meaning hypotheses for each word; when participants guessed incorrectly on a previous learning instance, they were at chance at selecting among alternatives during the

[3] For every trial for a given subject, one of the four competitor objects was randomly selected as the competitor using a random number generator. This procedure was used, rather than averaging all four competitors together, to make sure the Target and Competitor data matched in terms of data sample size and variance.

[4] Because proportion scores are not appropriate for linear models, such as ANOVAs, we first transformed the Target and Competitor proportions separately using an Empirical-Logit (e-logit) transform before taking the difference between the two (see, e.g., Barr, 2008). In addition, multi-level models were used rather than ANOVAs on subject and item means because of uneven N across conditions of Previous Accuracy.

[5] Means were estimated from the intercept values calculated using two multi-level models on the subsets of the data where Previous-Instance Accuracy was either Correct or Incorrect. In each case, the 'null model' was used, i.e., no fixed effects, but crossed random intercepts for Subjects and Items. *P*-values for estimated means in these model subsets were calculated by generating 10,000 Markov Chain Monte Carlo samples via the pvals.fnc function in R. All subsequent reporting of estimated means for Previous-Instance Accuracy subsets use this method.

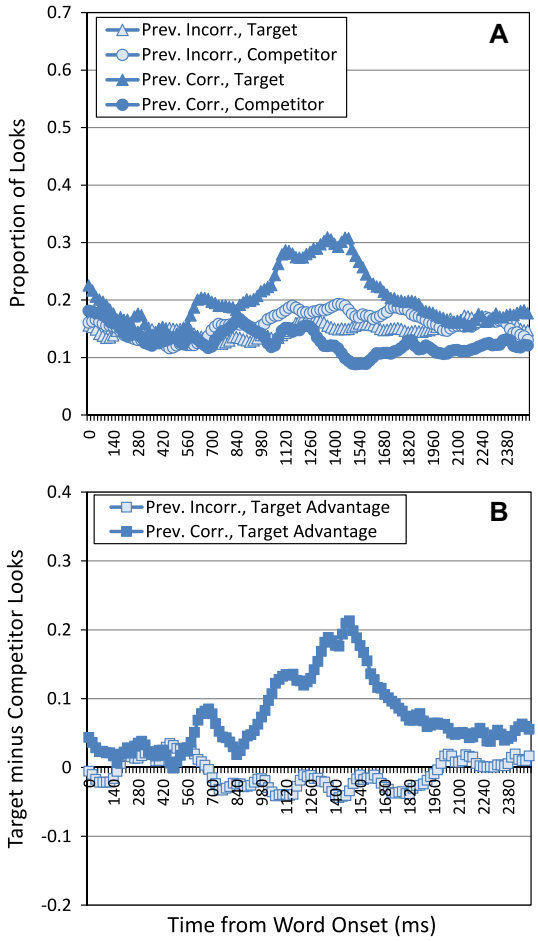

Fig. 3. (A) Average proportion of looks to the target referent (triangles) and a randomly selected competitor referent (circles) plotted over time from word onset. Dark filled symbols represent instances on which the participant had been correct on the previous instance. Light filled symbols represent instances on which the participant had been incorrect on the previous instance. (B) Target Advantage Scores (TASs): Proportion of looks to the target minus proportion of looks to the competitor. Smoothed Participant Means (Experiment 1).

next learning instance. This occurred even though the correct target referent was present during both learning instances.

However, it is possible that participants' general failure to recall referential alternatives was due to the high degree of referential ambiguity present during the previous learning instance: Participants in Experiment 1 would likely have to remember multiple alternative referents from the previous learning instance to achieve above chance performance following an incorrect trial. Although we have demonstrated elsewhere that high referential ambiguity is the norm in natural word learning environments (Gillette et al., 1999; Medina et al., 2011), we felt it prudent to reexamine the contingent learning patterns from Experiment 1 in a second experiment, in which referential ambiguity was markedly reduced. This was done by testing "High Informative" (HI) learning instances: those that contain only two referents as opposed to the four LI learning instances of Experiment 1 (see Fig. 4). After guessing incorrectly on a HI learning instance, participants need only remember one other referential alternative. Using the same accuracy-contingent analyses as above in Section 2.3.2, we can ask whether participants can recall this single (correct) referential alternative on the next learning instance.

3.2. Method

3.2.1. Participants

An additional 29 undergraduates participated. They had the same background as those individuals in Experiment 1.

3.2.2. Procedure

The procedure was the same as Experiment 1.

3.2.3. Stimuli and design

The stimuli were the same as in the previous experiment, with the following exceptions. On certain learning instances we reduced the number of referent images on the screen from five down to two, leaving just the correct target referent and one incorrect referential alternative. On these trials (henceforth HI learning instances) the two images appeared in positions to the left and the right of the central fixation, with position randomized. See Fig. 4 for an example. In the two-referent trials, each image was located approximately 6.7° from the central crosshair.

We created new experimental lists based on those used in Experiment 1. We preserved the order of words in the list, but we selected 8 out of the 12 items (*ball*, *bear*, *bird*, *book*, *car*, *door*, *fish*, and *hand*) to serve as target words – for these words, both 2-referent and 5-referent displays were used. Four nouns (*dog*, *cat*, *eye*, and *shoe*) were treated as filler words – for these words, only 5-object displays were used. We selected the target words such that two Target words were always followed by a Filler word, and thus Target words appeared as word numbers 1, 2, 4, 5, 7, 8, 10, and 11 within each 12-item block. Reverse lists preserved the Target–Target–Filler order, such that word order within block became 11–10–12–8–7–9–5–4–6–2–1–3.

We created two between-participant conditions: HI First and HI Middle. In the HI First lists, we changed the first learning instance of each Target item from a 5-alternative learning instance to a 2-alternative learning instance, by randomly dropping 3 of the incorrect referential alternatives. The result was that in the HI First condition, the first learning instance of each target word was a HI learning instance. All remaining learning instances were LI (Low Informative), five-alternative, instances. We then created the HI Middle lists from the HI First lists. We did this by moving the first round of learning (the first block) forward in the list by 2 rounds. Thus participants in the HI Middle condition encountered two rounds of LI learning instances (rounds 2 and 3 of the HI First condition) then a round of HI learning instances (round 1 from the HI First condition) followed by two more rounds of LI instances (rounds 4 and 5 from HI First). This was done to preserve the overall

Fig. 4. Example sequence of two learning trials for the word "zud", which meant 'bear'. The first learning trial (left) is a High Informative trial, in which only two alternatives are displayed; the second (right) is a low informative trial, in which five alternatives are displayed. Note that different examples of bears appear, that only one entity is labeled on each trial (i.e., only one nonsense word is heard), and that other learning trials for other words intervene.

word-referent statistics across lists when one ignores order. Fifteen participants were randomly assigned to the HI First condition and 14 to HI Middle.

3.3. Results and discussion

3.3.1. Learning curve

Fig. 5A plots average proportion of correct responses on target items over the five learning instances split by condition. Chance performance was .20 on all learning instances except for the first learning instance of the HI First condition and the third learning instance of the HI Middle condition – in both of these, chance was .50.

Rather than the gradual increase in average accuracy found in Experiment 1, we see that learning is occurring only after participants encountered a HI instance. In particular, in the HI Middle condition, learning instances before the HI item (Instances 1 and 2, henceforth called "A instances") had average accuracy near chance level (of .20) whereas learning instances after the HI item (Instances 4 and 5, henceforth "B instances") were above chance, around .40. In HI First, all four LI instances occurred after the HI instance, and indeed all means were above chance (again, around .40). For the HF condition, we can label Instances 2 and 3 as "A" instances and Instances 4 and 5 as "B" instances because, given the experimental design, they correspond to Learning Instances 1 & 2 and 4 & 5 respectively from the HI Middle condition. And indeed, a multi-level mixed model (Table 2) reveals a reliable interaction between Instance Type (A vs. B) and Condition (HF vs. HM), precisely because the accuracy of HM instances improved when going from A to B whereas the accuracy of HF instances began and remained high.

3.3.2. Accuracy-contingent clicking responses

To assess the learning mechanism responsible for the aggregate pattern reported in the previous section, we performed the following accuracy-contingent analysis, the results of which appear in Fig. 5B. We examined accuracy on the learning instance that immediately followed the HI instance (Instance 2 in HI first and Instance 4 in HI Middle), split by whether the participant had clicked on the correct or incorrect referent on the HI instance (e.g., split by whether the participant clicked on the bear or the door in the example in Fig. 4 above).

The pattern of responding was identical to the findings of Experiment 1, even though in the present experiment there were only two referential alternatives present on the preceding trial. After guessing incorrectly on a HI instance, participants randomly selected a referent, resulting in chance performance of .20. But, after guessing correctly on a HI instance, participants were well above chance. This

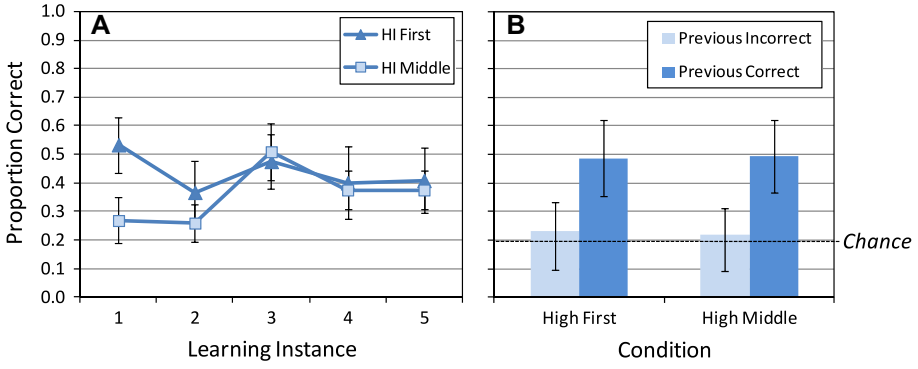

Fig. 5. Mean proportion of correct responses. Participant means. Error bars indicate ± 95% C.I. (A) As a function of learning instance and split by position of High Informative (HI) learning instance. (B) As a function of whether the participant had been correct or incorrect on the previous High Informative (HI) learning instance for that word. Data come from Instance 2 for HI First and Instance 4 for HI Middle. Number of participants contributing to bars from left to right: 15 out of 15, 15 out 15, 13 out of 14, 14 out of 14 (Experiment 2).

Table 2
Learning in the aggregate: Effects of Instance Type (A vs. B) and Condition (HF vs. HM) on accuracy (Experiment 2).

Effect	Estimate	S.E.	*z*-value	*p*-value
Intercept	−0.36	0.25	−1.41	.16
Instance Type (A vs. B)	−0.13	0.24	−0.54	.59
Condition (HF vs. HM)	−0.75	0.40	−1.88	.06
Interaction (Instance Type × Condition)	0.69	0.29	2.36	.02*

Note: Results of multi-level logit model with crossed random slopes and intercepts for both Participants and Items. The lmer function in *R* was used [syntax: Accuracy ~ 1 + InstanceType * Condition + (1 + InstanceType | Participant) + (1 + Condition + InstanceType | Item)]. Model was a better fit than a corresponding model that did not include interaction term as fixed effect ($\chi^2(1) = 5.23$, $p = .02$) and a marginally better fit than a model that had no fixed effects ($\chi^2(3) = 6.95$, $p = .07$). Instance Type A corresponds to Instances 1 and 2 in the HM condition and 2 and 3 in the HF condition, and Instance Type B corresponds to Instances 4 and 5 in both conditions.
* Statistically significant.

was true for both the HM and the HF condition. Multi-level logit modeling revealed only an effect of whether the participant had been correct on the previous trial, with no effects of or interaction with Condition (Table 3).

It is especially worth noting that the benefit of guessing correctly on a preceding HI instance is numerically quite similar to the benefit of guessing correctly on a preceding LI instance. In particular, compare Figs. 2B–5B. After guessing correctly on a LI trial, or a HI First instance, or even a HI middle instance, accuracy is within .46–.49 in all three cases, and well above the chance performance seen after guessing incorrectly on such trials. This suggests that all that matters is guessing correctly on a previous trial, not whether that previous trial had a high or low number of referential alternatives. Such a pattern is expected if participants retain only a single hypothesis about the meaning of the word, rather than store alternative hypotheses.

3.3.3. Eye movements

As in Experiment 1, it is possible that although participants' clicking behaviors indicate that participants only remember a single hypothesized meaning, their eye movements might show some implicit memory for the alternate hypothesis. Following the procedure used in Experiment 1, we examined this possibility by plotting the average proportion of looks to the Target and a randomly selected Competitor referent, from the onset of the word. Here we are plotting only those instances that immediately followed a HI learning instance (i.e., Instance 2 in HI First and Instance 4 in HI Middle).

As can be seen in Fig. 6, the eye movement patterns are similar to what was observed in Experiment 1 (compare to Fig. 3 above). Target looks exceeded Competitor looks only when the participant had been correct on the previous instance. When the participant had been incorrect on the previous learning instance, the average proportion of Target looks and Competitor looks were very similar. This again suggests that participants had no implicit memory that the previous learning instance included the Target referent. Instead participants recalled only the current working hypothesis about the word's meaning.

Table 3
Accuracy-contingent analysis. Best fitting model (Experiment 2).

Effect	Estimate	S.E.	*z*-value	*p*-value
Intercept	−1.20	0.25	−4.79	<.00001*
Accuracy on Prev. Inst. (Incorr. vs. Corr.)	1.16	0.30	3.90	.00001*

Note: Results of multi-level logit model with crossed intercepts for both Participants and Items. The lmer function in *R* was used [syntax: Accuracy ~ 1 + PreviousAccuracy + (1 | Participant) + (1 | Item)]. Model was a better fit than a model with fixed effects ($\chi^2(1) = 15.15$, $p < .001$). Models adding fixed effect of Condition or Condition × PreviousAccuracy interaction did not reliably improve fit. Adding random slopes to null model did not improve fit.
* Statistically significant.

These conclusions found support in multi-level linear modeling of the eye movement data using the same methods as described in the Results of Experiment 1, looking at the same four 500 ms windows. Using this method, it was found that Previous-Instance Accuracy was reliable in the third (1000–1499 ms) time window ($\chi^2(1) = 6.26$, $p < .01$) and fourth (1500–1999 ms) time window ($\chi^2(1) = 5.14$, $p = .02$). Moreover, in the third time window TAS was reliably positive when the participant had been correct on the previous trial (est. mean = 1.13, $SE = 0.46$, $t = 2.46$, $p = .04$), and marginally in the fourth time window (est. mean = 0.88, $SE = 0.38$, $t = 2.33$, $p = .08$), but not when the participant had been incorrect (third time window: est. mean = −0.17, $SE = 0.44$, $t = -0.38$, $p = .69$; fourth time window: est. mean = −0.16, $SE = 0.30$, $t = -0.55$, $p = .59$).

3.3.4. Does clicking matter?

It is possible that requiring a response on every learning instance may have influenced the kind of learning procedure used by the participants in the present experiments. To address this concern,

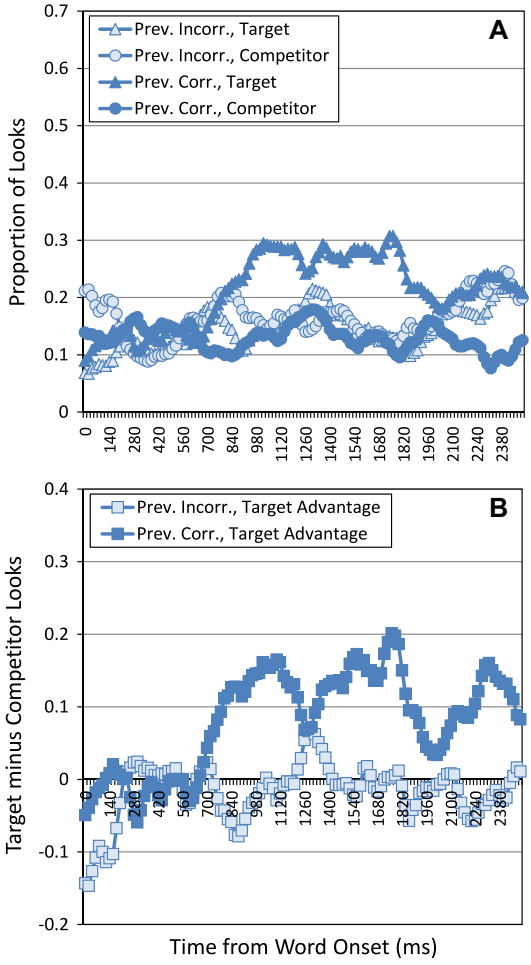

Fig. 6. (A) Average proportion of looks to the target referent (triangles) and a randomly selected competitor referent (circles) plotted over time from word onset. Dark filled symbols represent instances on which the participant had been correct on the previous instance. Light filled symbols represent instances on which the participant had been incorrect on the previous instance. Data was taken from the instance immediately following a High Informative (HI) learning instance (i.e., Instance 2 in HI First and Instance 4 in HI Middle). (B) Target Advantage Scores (TASs): Proportion of looks to the Target minus proportion of looks to the Competitor. Smoothed Participant Means (Experiment 2).

versions of Experiments 1 and 2 were also run in which participants ($N = 47$) were not given instructions to click on the objects until the fourth learning instance, in the conditions of HI Absent (Exp. 1), HI First (Exp. 2), and HI Middle (Exp. 2). Average clicking accuracies on the fourth learning instance in these 'delayed clicking' experiments were above chance and, importantly, quite similar to the fourth learning instances in Experiments 1 and 2. For HI Absent, accuracy was 32% (as compared to 31%); for HI First it was 37% (compared to 40%); and for HI Middle it was 43% (compared to 38%). Multilevel logit models like those used above found no effect of experiment (Clicking vs. Delayed Clicking) within any of these three conditions (HI Absent, HI First, or HI Middle) for the fourth learning instance, suggesting that clicking during instances 1–3 had no effect on learning. We therefore continue to employ clicking as a measure of word learning over time, in one final experiment that further reduces referential uncertainty.

4. Experiment 3

4.1. Introduction

Experiment 2 demonstrated that even when the number of referential alternatives on a previous learning instance is reduced from five to two, participants continue to use only their memory of the referent they had selected (or, more precisely, the current hypothesis of what the word means), even though memory demands were greatly reduced. Thus, simplifying the learning situation did not induce the learner to track multiple hypotheses. Here we explore this further, by another simplification of the learning situation. Recall from Experiment 2 that we reduced the referential alternatives on one learning instance (either the first or third learning instance among five). Here we ask what learning is like when multiple learning instances are highly informative: the sequence of five learning instances now will begin with either two, three, or four HI learning instances. We also include a fourth condition in which all five learning instances are HI. Using the contingent-response analyses, we can again ask if participants take advantage of the alternative (unselected) referent. Perhaps under these radically simplified learning conditions, participants switch to a learning strategy that tracks multiple hypotheses for each word.

4.2. Method

4.2.1. Participants

A total of 63 undergraduates participated and had the same background as the participants in Experiments 1 and 2.

4.2.2. Procedure

The procedure was the same as Experiments 1 and 2.

4.2.3. Stimuli and design

Four different experimental lists were created that reflected the number of HI instances to be provided to participants. In the 2-HI condition, participants received two HI learning instances follow by three LI instances. In the 3-HI, they received three HI and then two LI. In the 4-HI, they received four HI and then one LI. In the 5-HI, they received HI instances on all five learning instances.

These lists were created from the HI-First experimental list used in Experiment 2. For the 2-HI list, we changed all second learning instances on target trials from being 5-alternative to 2-alternative referent choices in the manner described above in Section 3.2.3 of Experiment 2. This new list was used to create the 3-HI list by reducing all third learning instances on target trials from 5- to 2-alternative, and so on for the 4-HI and 5-HI lists. Note that filler items remained LI throughout all five learning instances. Reverse lists were also created in which the order of trials within each learning instance block was reversed such that two Targets were always followed by a Filler, in the manner described in Experiment 2.

4.3. Results and discussion

4.3.1. Learning curve

Fig. 7A plots for each condition the average proportion of correct responses across the five learning instances. As seen in the figure, learning is very successful in the 5-HI condition; performance gradually climbs from chance performance (.50) on Instance 1 to .90 on Instance 5. In the other three conditions (2-HI, 3-HI and 4-HI), performance keeps pace with the 5-HI condition until the first LI instance is encountered. In no cases do participants drop back down to chance performance (.20) but rather are always above chance: .40 after 2-HIs, .57 after 3-His, and .67 after 4-HIs.

4.3.2. Accuracy-contingent clicking responses

The aggregate performance shown in Fig. 7A demonstrates that participants are clearly learning across multiple HI learning instances. Here we explore the mechanism underlying this learning via contingent analyses. Fig. 7B plots performance on these HI learning instances, split by whether the participant had been correct or incorrect on the immediately previous learning instance. This is plotted separately for Instances 2–5. Note that because we are only including HI instances, chance performance is .50 on all learning instances. Also, because of the experimental design, the amount of data contributing to means in the graph drops across learning instances, such that Instance 2 includes all four conditions, Instance 3 includes just 3-HI, 4-HI, and 5-HI; Instance 4 includes just 4-HI and 5-HI; Instance 5 includes just the 5-HI condition.

There are two striking aspects of the graph in Fig. 7B. First, even though the learning situation has been greatly simplified, performance remains stubbornly near chance immediately after an incorrect learning instance, regardless of which learning instance is examined (2–4). Second, performance is well above chance after a correct learning instance, and steadily improves across learning instances (from .78 correct on Instance 2 to about .93 correct on Instance 5). Such a pattern is consistent with the propose-but-verify learning account in that again, participants are not showing significant memory for alternative referents from previous learning instances. For example, a participant who incorrectly selects a door rather than a bear when hearing "zud" is very likely to be at chance on the next instance of "zud" even though the choice is now between, e.g., a bear and a shoe.

A multi-level logit model of these data (Table 4) supports these conclusions. Accuracy was modeled using fixed effects of Instance (2–5) and Previous Accuracy (Incorrect, Correct). As shown in the table, there was a positive effect on Accuracy for being Previously Correct and no effect of learning Instance,

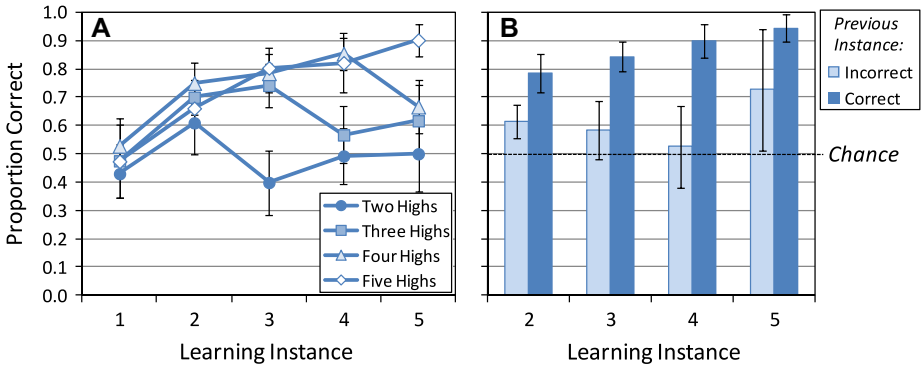

Fig. 7. Mean proportion of correct responses. Participant means. Error bars indicate ± 95% C.I. (A) As a function of learning instance and split by number of initial High Information (HI) learning instances: either two, three, four, or five HI instances. (B) As a function of whether the participant had been correct or incorrect on the previous High Informative (HI) learning instance for that word. Only HI trials were included (making chance performance .50). Number of subjects contributing to bars from left to right: Instance 2: 62 out 62, 62 out of 62; Instance 3: 46 out 47, 47 out of 47; Instance 4: 27 out of 32, 32 out of 32; Instance 5: 10 out of 16, 16 out of 16 (Experiment 3).

Table 4
Accuracy-contingent analysis. Best fitting model (Experiment 3).

Effect	Estimate	S.E.	*z*-value	*p*-value
Intercept	0.61	0.16	3.70	.0002*
Instance (2–5)	0.16	0.12	1.35	.17
Accuracy on Prev. Inst. (Incorr. vs. Corr.)	1.14	0.15	7.42	<.00001*
Interaction (Instance × Prev. Acc.)	0.41	0.16	2.52	.01*

Note: Results of multi-level logit model with crossed intercepts for both Participants and Items, plus a random slope for the effect of Instance with Participants. The lmer function in *R* was used [syntax: Accuracy ~ 1 + Instance * PreviousAccuracy + (1 + Instance|Participant) + (1|Item)]. Models with more complex random slope designs suffered from correlated random factors; however, an analysis using random slopes for both fixed effects within both Participant and Items yielded the same significance patterns as reported here. The above model was a better fit than a corresponding model that did not include interaction term as fixed effect ($\chi^2(1) = 6.05$, $p = .01$) as well as a corresponding model without any fixed effects ($\chi^2(3) = 80.3$, $p < .001$). Instance variable was centered.
* Statistically significant.

but a reliable interaction between the two factors, which presumably arose because Accuracy increases with Instance only when the previous learning instance was correct (see Fig. 7B).

Note that the error bars in Fig. 7B, which represent 95% confidence intervals around participant means, offer some (albeit weak) evidence that participants may be recalling the rejected referential alternative from a previous learning instance. Specifically, average accuracy on Learning Instances 2 and 5 are slightly above chance when the participant had previously been incorrect for that word (the lighter colored bars, whose error bars do not cross over the .50 chance level). We suspect that these effects are fragile, if not completely spurious, because we do not observe similar above-chance performance in Instances 3 or 4. Accounts that propose that multiple hypotheses are tracked across learning instances would predict above chance performance across all four of these learning instances. Moreover, although the error bar does not touch the .50 chance level in the Previous Incorrect condition of Instance 5, this above-chance performance was only marginally significant in a *t*-test of subject means ($t(9) = 2.07$, $p = .07$) and not significant in a *t*-test of item means ($t(6) = 1.26$, $p = .26$).

The current experimental design also allows us to test a further prediction of the propose-but-verify learning procedure. In particular, we can perform a slightly more complex contingent analysis, asking what drives response accuracy more: accuracy on the immediately preceding learning instance (one back), or accuracy on the learning instance that occurred two back? If learners are carrying only a single hypothesis across learning instances, performance on an instance that was two-back should have no effect on the current learning instance; only the previous learning instance (one back) should matter. This further dividing of the data – into four possible sequences: (1) Incorrect, Incorrect; (2) Correct, Incorrect; (3) Incorrect, Correct; (4) Correct, Correct – is possible here because with only two-alternative choices on each learning instance data are more evenly distributed amongst these four possible outcomes (roughly 25% in each).[6] Fig. 8 shows this analysis, plotting accuracy on the third learning instance as a function of the accuracy pattern on the previous two learning instances. Data were included from only the 3-HI, 4-HI, and 5-HI conditions, resulting in chance performance being .50.

Strikingly, the only factor that appears to influence accuracy on learning Instance 3 is accuracy on learning Instance 2. No effect of learning Instance 1 is apparent, nor is an interaction between these two factors. These observations were supported in a multi-level logit model (Table 5), which reveals a reliable effect of Instance 2 on Instance 3 Accuracy, but no effect of Instance 1 on Instance 3. Adding an interaction term did not improve the fit of this model ($\chi^2(1) = 0.69$, $p = .41$).

4.3.3. Eye movements

Following the procedure used in Experiments 1 and 2, we examined whether participants considered the alternative meanings by plotting the average proportion of looks to the Target against a Com-

[6] Five-alternative choices (used in Experiments 1 and 2) did not provide enough data in each sequence to make this additional analysis possible.

Table 5
Accuracy-contingent analysis of HI Instance 3 (Experiment 3).

Effect	Estimate	S.E.	*z*-value	*p*-value
Intercept	0.45	0.27	1.66	.10
Accuracy on Instance 1 (Incorr. vs. Corr.)	0.24	0.29	0.83	.41
Accuracy on Instance 2 (Incorr. vs. Corr.)	1.20	0.27	4.39	.00001*

Note: Results of multi-level logit model with crossed intercepts for both Participants and Items, plus a random slope for the effect of Accuracy on Instance 1 with Participants. The lmer function in R was used [syntax: Accuracy ~ 1 + AccuracyInst1 + AccuracyInst2 + (1 + AccuracyInst1 | Participant) + (1 | Item)]. Models with more complex random slope designs suffered from correlated random factors; however, an analysis using random slopes for both fixed effects within both Participant and Items yielded the same significance patterns as reported here. The above model was a better fit than a corresponding model without any fixed effects ($\chi^2(2) = 20.9$, $p < .001$). Adding an interaction term to the above model did not reliably improve the fit ($\chi^2(1) = 0.69$, $p = .41$).
* Statistically significant.

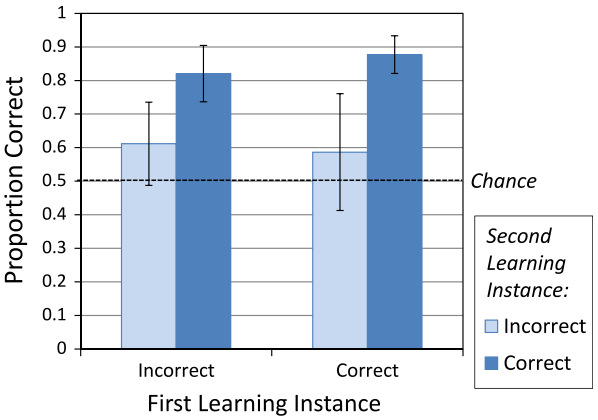

Fig. 8. Mean proportion of correct responses on HI learning Instance 3, as a function of whether the participant had been correct or incorrect on the first and second learning instances for that word. Only HI trials were included (making chance performance .50). Error bars indicate ± 95% C.I. Number of subjects contributing to bars from left to right: 39 out of 47, 45 out of 47, 26 out of 47, 45 out of 47 (Experiment 3).

petitor referent, from the onset of the word, split by whether participants were Correct or Incorrect on the previous instance (see Fig. 9). Here we plot only HI instances that immediately followed another HI instance – that is, Instance 2 from the 2-HI condition, Instances 2 and 3 from the 3-HI, Instances 2, 3, and 4 from the 4-HI, and Instances 2, 3, 4, and 5 from 5-HI.

As can be seen in Fig. 9, the average proportion of looks to the Target and Competitor were in general higher than what was observed in previous experiments (compare to Figs. 3 and 6 above) – a difference that is expected because trials in the present experiment had only two potential referents on the screen rather than the five potential referents used in the previous experiments. Under these very simplified referential conditions – and only these – do we observe any hint that participants are implicitly recalling the alternative referent from the previous learning instance. Target looks greatly exceeded Competitor looks when the participant had been correct on the previous instance, but Target looks also exceeded Competitor looks when the participant had been incorrect on the previous instance, albeit to a lesser extent. Thus there is some implicit memory of the previous referential alternative, although it has very little impact on overt referential choices.

These conclusions found support in multi-level linear modeling of the eye movement data using the same methods as described in the Results of Experiment 1 and 2, looking at the same four 500 ms time windows. Using this method, it was found that Previous-Instance Accuracy was reliable in the 0–499 ms time window ($\chi^2(1) = 5.74$, $p = .02$), 500–999 ms time window ($\chi^2(1) = 24.9$, $p < .001$),

1000–1499 ms time window ($\chi^2(1) = 20.4$, $p < .001$), and 1500–1999 ms time window ($\chi^2(1) = 4.34$, $p = .04$). For those times when the participant had been correct on the previous instance, TAS was found to be reliably positive in all of these time windows: 0–499 ms (est. mean = 0.53, $SE = 0.16$, $t = 3.32$, $p = .02$); 500–999 ms (est. mean = 2.30, $SE = 0.38$, $t = 6.00$, $p < .001$); 1000–1499 ms (est. mean = 2.47, $SE = 0.34$, $t = 7.30$, $p < .001$); 1500–1999 ms (est. mean = 1.35, $SE = 0.26$, $t = 5.16$, $p < .001$). But unlike the first two experiments, here TAS also became reliably positive for trials on which the participant had been incorrect on the previous instance, at least for the 1000–1499 ms window (est. mean = 1.07, $SE = 0.48$, $t = 2.23$, $p = .05$) and the 1500–1999 ms window (est. mean = 0.71, $SE = 0.33$, $t = 2.13$, $p = .05$). This finding suggests participants did have some implicit memory of the referential alternative present on the previous learning instance.

4.3.4. Summary of Experiment 3

Even under a greatly simplified learning situation (far simpler than that faced by a person learning a language from the natural referent world), the response patterns of participants continue to be strikingly consistent with the propose-but-verify learning procedure. The clicking response patterns suggest that participants had a strong tendency to recall and use only a single hypothesized meaning of the word, as derived from the immediately preceding learning instance, and not the alternative referent from a previous learning instance. Though the eye-movement findings suggest some implicit memory of the alternative referent, this does not appear to influence overt responses.

5. Simulation of the results from Experiments 1–3

Although we have argued that our findings are predicted by the propose-but-verify account, it is worth establishing that a computer-implemented version of the account can simultaneously capture the aggregate learning effects (as seen in Figs. 2A, 5A, and 7A) as well as the response-contingent patterns (as seen in Figs. 2B, 5B and 7B).[7]

5.1. Simulation method

We ran computer simulations of the propose-but-verify account using a simple algorithm that can be described as:

1. Begin by guessing at chance.
2. On any additional occurrence of a word, remember the previous guess with some probability α.
3. If the remembered guess is present in the current referent set (i.e., confirmed), increase α and select the referent; otherwise select a referent at random.

Thus α is the only free parameter in the model. We can estimate α from the experimental data as follows. In the Experiment 1 design (i.e., five LI instances), the chance of selecting a particular referent is .20. Consider the situation in which the participant has selected the correct referent on the immediately preceding learning instance. Under this situation, correctly selecting the target again can come about from one of two possible scenarios: (1) the participant successfully recalled the previous response and thus selected it again, or (2) the participant failed to recall the previous response but then happened to select the target referent at random. The first scenario occurs with a probability of α whereas the second scenario occurs with a probability of $(1 - \alpha)$ times chance. Thus, after selecting the target correctly on a learning instance, accuracy on the next learning instance is:

$$\text{accuracy} = \alpha + (1 - \alpha) * \text{chance}$$

By solving for α we can estimate the probability of recalling a previous response:

[7] Ongoing work (Stevens, Yang, Trueswell, & Gleitman, 2012) uses a more sophisticated version of this model to attempt to capture the experimental findings of word learning while at the same time to compete, or in some cases outperform, wider-coverage models of word learning derived from corpora of child-directed speech.

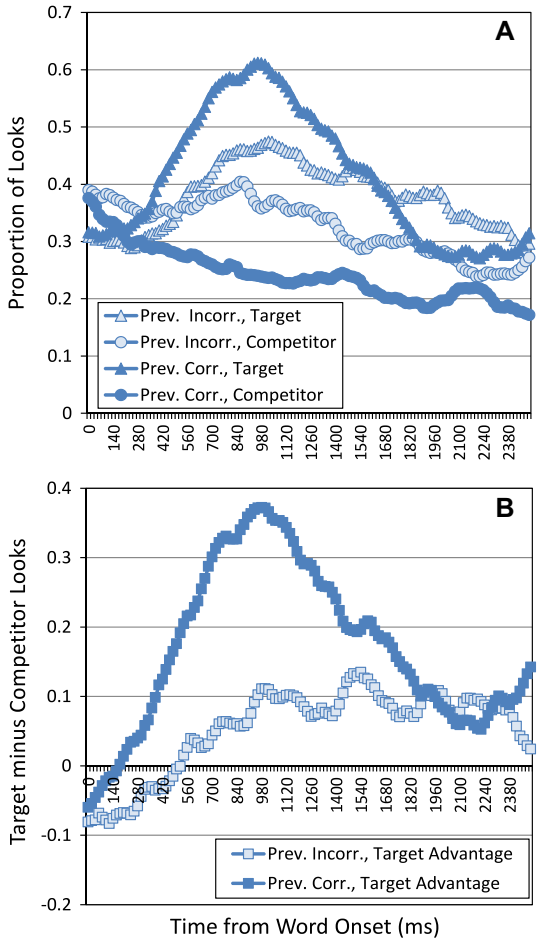

Fig. 9. (A) Average proportion of looks to the target referent (triangles) and the competitor referent (circles) plotted over time from word onset. Dark filled symbols represent instances on which the participant had been correct on the previous instance. Light filled symbols represent instances on which the participant had been incorrect on the previous instance. Data was taken from only the High Informative (HI) Instances that also followed a HI Instance. (B) Target Advantage Scores (TASs): Proportion of looks to the Target minus proportion of looks to the Competitor. Smoothed Participant Means (Experiment 3).

$$\alpha = (\text{accuracy} - \text{chance})/(1 - \text{chance})$$

For simulations of Experiments 1 and 2, we set the initial value of α based on the average accuracy of those LI learning instances where participants had been correct for the first time on the immediately preceding learning instance, i.e., correct on the previous learning instance but not correct on earlier learning instances (a total of 331 trials from Experiments 1 and 2). Under these conditions, accuracy was 0.41,[8] making $\alpha = 0.26$. Because there were only 13 HI trials for which the participant had been correct on the previous trial, we made $\alpha = 0.26$ on HI trials as well.

[8] This accuracy score was quite similar when the previous learning instance had been a LI (.40) or a HI (.45), which did not differ in a two-tailed t-test on subject means ($p > .76$). This lack of a difference is consistent with the notion that the only thing recalled from a previous learning instance was the hypothesized meaning and not alternative hypotheses (see Medina et al., 2011 for a similar finding).

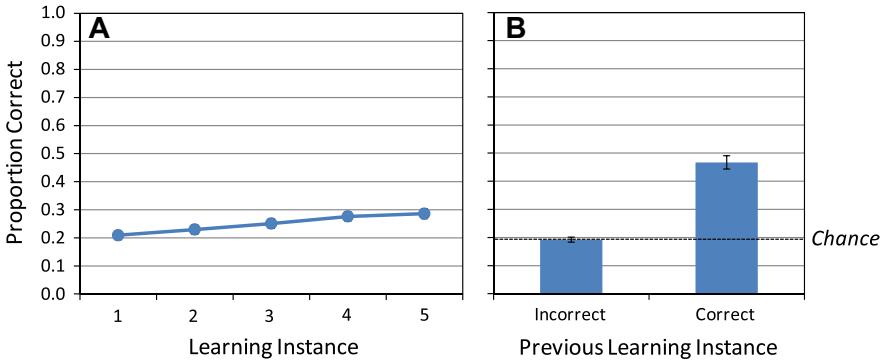

Fig. 10. Mean proportion of correct responses from 200 simulated subjects. Error bars indicate ± 95% C.I. (A) As a function of learning instance. (B) As a function of whether the participant had been correct or incorrect on the previous learning instance for that word (simulation of Experiment 1).

Once a hypothesis has been confirmed, we would expect α to increase dramatically, i.e., participants should remember a confirmed hypothesis at a much greater rate. Indeed, after accurately responding to two learning instances in a row, participants went on to have an accuracy of 0.77 on the next instance in that sequence (a total of 235 trials), making $\alpha = 0.71$. Thus, once a hypothesis was confirmed, we raised α from 0.26 to 0.71 in simulations of Experiments 1 and 2.

The same procedure was used to set α for Experiment 3, except we used HI trials rather than LI trials to determine α. That is, we used the average accuracy of those HI trials that followed a first correct guess to determine the initial α (here $\alpha = 0.60$, 391 trials), and the average accuracy of HI trials after a confirmed guess for the increased α ($\alpha = 0.81$, 399 trials). One should expect α to be higher in Experiment 3 than in Experiments 1 and 2 because the reduced number of distractors on a current learning instance (2 rather than 5) should make it easier to remember the current hypothesis for the word. It was not possible to set a different α for LI trials in Experiment 3 because of the relatively few number of trials on which a LI trial was preceded by a trial that was the first correct item, i.e., the first confirmed item. Ideally, one would want to determine empirically the relationship between the probability of remembering a previous response (α) and the number of referential distractors present on a current trial, but this would require a more extensive parametric manipulation of the referent set size than was done here in the present experiments.

We ran 200 simulations (200 simulated subjects) per condition in each experiment, with 100 simulated subjects assigned to each forward- and reverse-ordered stimulus list. Thus 200 simulated subjects 'participated' in Experiment 1, 200 each in the HI First and HI Middle conditions of Experiment 2, and 200 in each of the four conditions of Experiment 3. The accuracy for each response was recorded, and subject means and 95% confidence intervals were computed.

5.2. Simulation results

As can be seen in Fig. 10A, the 200 simulated subjects for Experiment 1 performed strikingly similar to the actual subjects (compare to Fig. 2A); the numerical values of each mean were almost identical to the actual subjects and fell within the ±95% confidence intervals of actual subject performance. And the same contingency effect was observed of being at chance if the simulated subject had been incorrect on the immediately preceding learning instance (Fig. 10B, compare to Fig. 2B). Finally, a multi-level model like the one found in Table 1 generated the same reliable effect of learning instance for the simulated data (est. slope = 0.14, $z = 7.61$, $p < .01$).

As can be seen in Fig. 11A, the simulations also captured the general patterns observed in Experiment 2 (compare to Fig. 5A). In particular, just like actual subjects, the simulated subjects showed fairly flat but above-chance performance on all LI trials after encountering a HI learning instance. It is true that the actual subjects out-performed the simulated subjects slightly, but the same overall

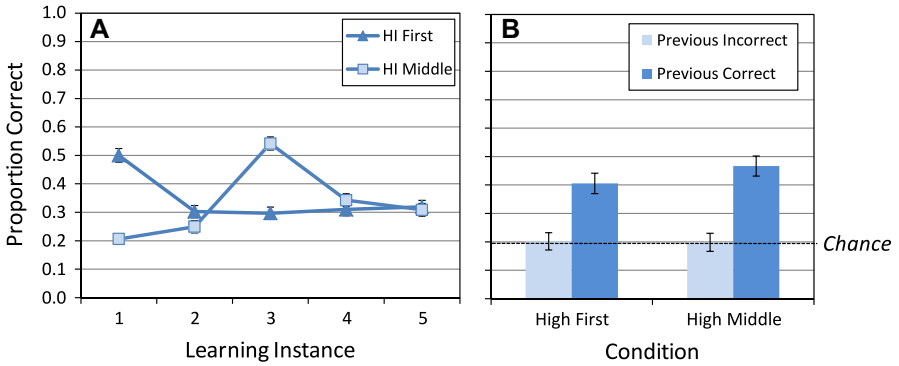

Fig. 11. Mean proportion of correct responses from 200 simulated subjects per condition. Error bars indicate ± 95% C.I. (A) As a function of learning instance and split by position of High Informative (HI) learning instance. (B) As a function of whether the participant had been correct or incorrect on the previous High Informative (HI) learning instance for that word (simulation of Experiment 2).

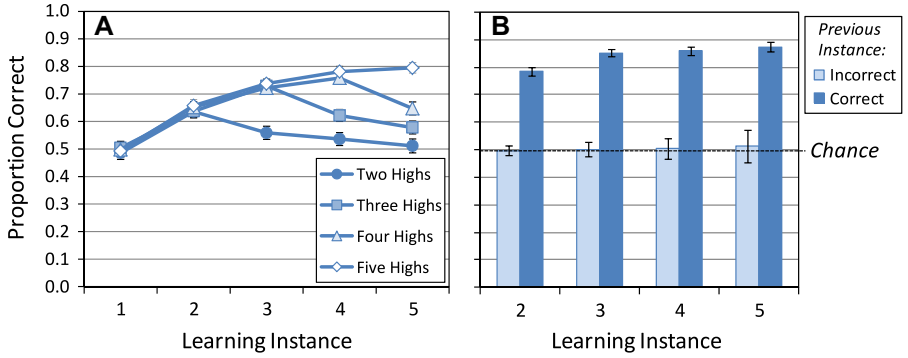

Fig. 12. Mean proportion of correct responses from 200 simulated subjects per condition. Error bars indicate ± 95% C.I. (A) As a function of learning instance and split by number of initial High Information (HI) learning instances: either two, three, four, or five HI instances. (B) As a function of whether the participant had been correct or incorrect on the previous High Informative (HI) learning instance for that word. Only HI trials were included (making chance performance .50) (simulation of Experiment 3).

patterns were observed. The contingency analyses of the simulated data also generated patterns quite like the actual subjects (compare Fig. 11B to Fig. 5B); here performance was well within the 95% Confidence Intervals of actual subjects. A multi-level model like the one found in Table 2 generated the same significance patterns on simulated subjects: a reliable effect of Condition (est. slope = 0.41, $z = -5.60$, $p < .01$) which interacted with Instance Type (est. slope = 0.52, $z = 6.42$, $p < .01$). In the same way, a multi-level model like the one found in Table 3 for actual subjects generated the reliable effect of Accuracy on Previous Instance (est. slope = 1.24, $z = 15.03$, $p < .01$).

Finally, Fig. 12A shows that the simulations also captured the general patterns observed in Experiment 3 (compare to Fig. 7A). Just like actual subjects, accuracy increased across a sequence of 5 HI learning instances (from 50% to 80% correct). And, like actual subjects, the simulated subjects' performance also dropped off once a LI trial was encountered in a sequence, but always remained well above chance. There was one notable difference between the simulated and actual data: Actual subjects tended to outperform simulated subjects on HI trials late in the sequence (i.e., instances 4 and 5). We believe this was because our learning algorithm did not increase the probability of remembering (i.e., α) even further when the hypothesis was confirmed for a second or third time; we simply kept it at the same value of 0.81. Many of these later correct responses from actual subjects were from trials

when the subject had been correct on at least two preceding instances, presumably making it even easier to remember the correct hypothesis. The contingency analyses on the simulated data of Experiment 3 (Fig. 12B) also generated patterns similar to the actual subjects (Fig. 7B). Here the simulated subjects are exactly at chance after guessing incorrectly on a previous learning instance, whereas the actual subjects were sometimes slightly above chance, albeit inconsistently so. Finally, a multi-level model like the one found in Table 4 generated the same significance patterns with simulated subjects: a reliable effect of Previously Correct (est. slope = 1.64, $z = 42.84$, $p < .01$) which interacted with Instance (est. slope = 0.16, $z = 4.13$, $p < .01$).

In sum, the resulting behavior of a very simple propose-but-verify model, with just one free parameter (α, the probability of remembering), captured both the average learning curves and the contingency patterns found in Experiments 1–3.

6. General discussion

6.1. Summary and key observations

In three studies of cross-situational word learning, participants rarely, if ever, retained multiple alternative hypotheses to a word's meaning across learning instances, even under conditions of low referential ambiguity. When participants guessed incorrectly on a learning instance, they were at chance when selecting among referents on the very next learning instance for that word, even though the target referent was present on both learning instances. This occurred under conditions where both learning instances contained five potential referents (Experiment 1, Fig. 2B) and where the first of the two learning instances contained only two potential referents (Experiment 2, Fig. 5B). The implicit measure, of tracking participants' eye movements to alternative referents, yielded a strongly supportive result: There were no indications in the eye movement data that participants remembered the referential alternative from the immediately preceding learning instance (Figs. 3 and 6). In Experiment 3 we reduced the learning task to its barest (and admittedly least realistic) variant by presenting participants with only two alternative referents to choose between on all relevant trials. Even in these least taxing circumstances, participant choices across trials were sensitive to the individual's hypothesis from the immediately preceding trial (Figs. 7B and 8), though in this one case there was a fragile suggestion in the eye-tracking data of memory for the competing referent in the preceding trial (Fig. 9). Finally, a simple computer simulation of the propose-but-verify procedure quite precisely captured the response patterns of our participants (see Figs. 10–12).

6.2. The present findings within the literature of word learning

Our results, and the results from Medina et al. (2011), are largely consistent with findings on vocabulary learning that have accumulated during the past three decades of research and theoretical commentary. Notably in this regard, both children and adults exhibit a striking ability to seize upon the correct meaning of a word upon first encounter, especially when the situational context is highly informative and constraining (e.g., Carey, 1978; Carey & Bartlett, 1978; Heibeck & Markman, 1987) with children sometimes using additional information to constrain meaning (e.g., Booth & Waxman, 2002; Fisher, Gertner, Scott, & Yuan, 2008; Gropen, Pinker, Hollander, & Goldberg, 1991; Nappa et al., 2009; Soja, Carey, & Spelke, 1991). Our results exhibit a similar pattern, but speak also to how such a fast mapping procedure works under less informative situations across multiple learning instances. Participants seize upon (or even stumble upon) a single hypothesized meaning and bring this forward for confirmation, or rejection, on the very next learning instance. The situational statistics make it much more likely that a correct hypothesis, when selected, will be confirmed when the word is heard again.

Yet, the results also seem to raise a paradox. They show that one-trial learning with cross-trial verification is a slow and laborious way of building a lexicon. Learning from sequences of the typical, relatively low informative learning instances found in natural contexts (Medina et al., 2011) or our artificially generated contexts with high referential uncertainty (Experiment 1 above), yield very slow

learning patterns in the aggregate (e.g., Fig. 2A). This appears to contradict what is known about child vocabulary growth, where children are estimated to be acquiring close to a word every couple of waking hours throughout the toddler and school years, with striking referential success (Bloom, 2002; Carey, 1978).

This apparent paradox falls away however when we consider the kind of contextual evidence offered to the participants here and in Medina et al. (2011) to perform cross-situational confirmation. In Medina et al. (2011), we stripped away all cues to a word's meaning, except for the buzzing-blooming visual referent world of the child; recall that Medina et al. eliminated all linguistic cues to word meaning by muting the videos, and providing just a single nonsense word at the occurrence of the 'mystery word'. This 'simulates' the contextual evidence that is available to the very early language learner who has not yet constructed a database of linguistic knowledge (of the meanings of other words, sentential syntax, etc.) – an assumption that has been given strong empirical support from other studies using the human simulation paradigm (Gillette et al., 1999; Snedeker & Gleitman, 2004). Likewise, in the present studies we have given adult learners a referent world to work from, but not much more, and have made it, appropriately so, referentially underdetermined (especially in Experiment 1).

Thus the slow growth of vocabulary is understandable if we are 'simulating' here the earliest stages of word learning, where a child knows just a few or even no additional words. Indeed, infant word learning appears to be a slow laborious process during this initial stage. For instance, the McArthur CDI infant scale of parent self-report routinely finds average production vocabulary going from 0 or 1 word at 8 months to just 20–25 words six months later at 14 months (e.g., Bates, Dale, & Thal, 1995), with measures of comprehension vocabulary also moving sluggishly from about 35 words to 140 words over this same six month period. Laboratory studies of comprehension show a similar pattern, starting with a small set of interpretable vocabulary items at six months moving to a slightly larger, but still quite small, set at 12–14 months (e.g., Bergelson & Swingley, 2012, and references therein).

It is only at 14-months when vocabulary growth begins its sharp climb, reaching about 12,000 words by age 6 years (e.g., Anglin, 1993; see also Bloom, 2002; Snedeker, 2009). What accounts for this rise? As we have argued elsewhere (Gillette et al., 1999; Gleitman, 1990), the acquisition of syntax and other linguistic knowledge by toddlers and young preschoolers during this time period provides a rich database of additional constraints that permit the learning of many additional words, both referential and non-referential. Indeed, these linguistic sources of evidence, when combined with the referent world, have been found to almost over-determine the meaning of many words (Gillette et al., 1999; Snedeker & Gleitman, 2004), dramatically increasing the rate of highly informative learning instances, which in turn, trigger additional sudden insightful learning on an item-by-item basis. Although we have not presented direct evidence here, it is quite plausible that a propose-but-verify word learning procedure is at work all along the course of word learning throughout most of the lifecycle (for word learning continues at a very high rate at least through the early years of adulthood, see Bloom, 2002). But as linguistic sophistication proceeds, the learner increasingly turns many low informative learning instances into highly informative ones by adding distributional, syntactic, and pragmatic information to that supplied by the observed world (Gleitman et al., 2005).

Plausibility aside, however, it is important to consider why and how the results just adduced seem so sharply variant with recent findings on statistical language learning. One could argue that statistical learning of vocabulary looks plausible too, perhaps more so than sudden, relatively insightful, learning. After all, there is no doubt that humans can track the statistical regularities associated with perceptual input, including input that is linguistic in nature (e.g., Aslin & Newport, 2008; Gebhart, Newport, & Aslin, 2009; Gómez & Gerken, 2000; Saffran et al., 1996). Moreover adults, when comprehending their native language, show exquisite sensitivity to the frequency of alternative structures and alternative meanings associated with ambiguous words (e.g., Duffy, Morris, & Rayner, 1988; Simpson & Burgess, 1985; Trueswell, 1996). Yet we have shown here that the patterns of cross-situational word learning tend not to show a similar probabilistic profile, even under fairly low degrees of referential uncertainty. And we have obtained similar results when using highly naturalistic stimuli (namely, the actual usage of parents speaking to 12- to 15-month olds as presented in HSP) that contain a high degree of referential uncertainty akin to what is faced by the child language learner (Medina et al., 2011).

There are two major issues that seem to lie behind these disparities of effect and implied theory. The first is simply methodological, having to do with how past data have been analyzed and reported, specifically, whether the evolving learning pattern within an individual, rather than just the aggregated end-state of learning, is assessed. The second potential disparity is more substantive. It arises from the artificiality of the stimulus situation that obtains in relevant laboratory studies – including our own as reported in the present paper – that purport to link with the *in vivo* facts about child vocabulary growth. We now take up these two issues in turn here.

6.2.1. The cross-trial evolution of vocabulary learning

As emphasized throughout the present paper, our ambition has been to measure participants' evolving lexical representations across exposures to new words in context, rather than solely to record final attainment. This is why we assessed our participants' conjectures after each learning trial, using both an explicit measure (their stated conjecture) and an implicit one (the movement of their eyes across the visually-present alternatives). The statistical word learning literature has, instead, typically reported solely or primarily on the presumed end point of learning by seeing whether participants are, as a group, above chance on guessing after all learning trials have been presented (Frank et al., 2009; Frank et al., in press; Xu & Tenenbaum, 2007; Yu & L. Smith, 2007).

These aggregate measures of final attainment leave unmeasured the participants' behavior on interim trials, simply presupposing that there is gradual (cross-trial) improvement as alternative conjectures rise or fall in probability as a function of changing cross-trial co-occurrence statistics. But this is not clear at all. If learning is characteristically abrupt, due to sudden insight on a single learning instance, and if these insights happen on different trials, across individuals, average performance will look gradual over that period of time.

This important methodological-analytic point has been noted again and again in the learning literature for at least the last half-century (see also Gillette et al., 1999, for evidence and discussion in the context of observational word learning). For example, Rock (1957) and Bower (1961) offered evidence that paired-associate learning, which in the aggregate looks like gradual learning, may instead produce a correct response "in an all-or-none fashion, and that prior to this conditioning event the participant guesses responses at random to an unlearned item" (p. 255, Bower, 1961). Likewise, Rock (1957) concluded from his own work that "in the classical multiple-item learning situation, associations are formed in one trial" (p. 190). And, as noted in Gallistel, Balsam, and Fairhurst (2004), a large number of other early memory researchers were well aware that average performance across participants can generate misconceptions about how learning is operating within the individual, citing, e.g., the work of Lashley (1942) and Estes (1956, 2002).

Indeed, Gallistel et al. (2004) show in a series of learning experiments that "in most subjects, in most paradigms, the transition from a low level of responding to an asymptotic level is abrupt" (p. 13124). The act of averaging across individuals who make this transition at different points in a learning sequence generates a gradual learning profile, exactly like that shown in Fig. 2A of the present paper. And our conditional analyses (e.g., Fig. 2B) suggest that cross-situational word learning is yet another example of this phenomenon. These outcomes suggest that the kind of "statistical learning" reported recently in the language learning literature actually belongs within this class as well, learning that is abrupt and solidified by confirmation.

It is worth highlighting, however, that a one-trial learning account of paired associative learning is a minority position in the current learning literature, yet it is unclear whether there is convincing experimental evidence to reject it in favor of a 'gradualist' learning position. This observation was most recently made in a historical review by Roediger and Arnold (2012), entitled "The One-Trial Learning Controversy and its Aftermath: Remembering Rock (1957)", which focused on how Rock's (1957) evidence for one-trial learning was dismissed, albeit with less than convincing evidence from opponents. As stated by the authors:

> Rock's conclusions rocked the world of verbal learning, because all theories followed a gradualist assumption. However, Estes (1960) published research that led him to the same conclusion shortly thereafter. Our paper. . .discusses how the verbal learning establishment rose up to smite down these new ideas, with particular ferocity directed at Rock. Echoing G.A. Miller (1963), we conclude

> with a note of sympathy for Rock's and Estes' positions and muse about why their work was so summarily dismissed. The important question they raised – the nature of how associations are learned – remains unanswered. (p. 2, Roediger & Arnold, 2012)

A one-trial learning account fits quite well with the fast-mapping literature on vocabulary learning that we have cited above (e.g., Heibeck & Markman, 1987). In this work, which tends not to focus on cross-situational evidence gathering, but rather the evidence present on a single learning instance, it has been found that word learning characteristically happens on just one trial. Analogous fast-mapping is also observed in other tasks that do not involve word learning (e.g., Markson & Bloom, 1997). This latter point comports well with the observations that sudden learning is more the norm than exception across a range of paradigms, learning problems, and participant groups, including other species (Gallistel et al., 2004).

One-trial learning does conflict with the views expressed by most researchers currently studying cross-situational learning, who largely adopt a gradualist position in which multiple associations between a word and its possible referents are tracked over time eventually converging on a single, most likely association. We reiterate here that the particular findings from this literature (e.g., Yu & L. Smith, 2007) are not at odds with our propose-but-verify account, as those findings come from final tests of learning rather than unfolding learning patterns. And there is no reason to suspect that a propose-but-verify learning procedure could not capture the final attainment patterns observed in that work.[9] Indeed, our experiments generate quite similar aggregate results: On average, people are more accurate on learning instance five than they were on learning instance one. And like Yu and L. Smith's findings, participants on instance five perform better after a series of highly informative trials as compared to a series of low informative trials. But these pooled scores are of little interest here because they leave unaddressed the learning mechanism that generated them.

Nevertheless some may wish to examine the differences between the present methods and those used by Yu and L. Smith (2007), the most well known of these cross-situational studies. Unlike the present method, Yu and L. Smith offered labels for visually presented referents on every trial. Labeling all referents no doubt speeds learning as it introduces the opportunity for reduction of referential uncertainty through a strategy of mutual-exclusivity (i.e., "moop" is confirmed by the presence of a bear, so "mipen" must be referring to the door). But this is an orthogonal issue to the one examined here, as it does not speak to how "moop" was learned in the first place. Second, one might argue that Yu and L. Smith's instructions were vague enough to encourage one-to-many mappings between word and referents. They do indeed say that subjects "were not told that there was one referent per word." Yet, they go on to say subjects "were told that multiple words and pictures would co-occur on each trial and that their task was to figure out across trials which word went with which picture" (p. 416, Yu & L. Smith, 2007). Thus the instructions implied that each word had a unique meaning. It is also true that Yu and L. Smith did not use carrier phrases such as "Look! A ___ !", as done in the present studies; we suppose the absence of these phrases could encourage the listener to infer that the words in the study were not referential in nature and instead, e.g., exclamatory (Fennell & Waxman, 2010; Fulkerson & Waxman, 2007; Namy & Waxman, 2000), but if true, this would undermine the claims of Yu and L. Smith that they were studying word referent-pairings under conditions of referential uncertainty. Finally, the passive nature of Yu and L. Smith's task, of not selecting referents on each trial, could have encouraged multiple referent tracking. Yet, our own comparisons of clicking and not clicking on referents suggest that this has little or no effect on learning (see Section 3.3.4 above).

[9] Very recently Yu and L. Smith (2012) have tested several learning models against their original (2007) data. Though, as they report, various versions of the models they investigated fulfilled different desiderata, the single-hypothesis-testing model, one much like our own, fit their human data most closely though, to be sure, other models could be adjusted in certain parameters so as to come close to the performance of this one. It of course remains true that in this work Yu and L. Smith had only final attainment data, i.e., they did not have any trial-by-trial data against which to test the evolution of learning under their models. This means that a central property of our model – the "verify" trial that succeeds the "propose" trial – cannot be evaluated there. Nonetheless, the new studies are a welcome further documentation of how single hypothesis machinery can and does work.

6.2.2. The world is so full of a number of things

Psychology through its history has benefited from examining radically simplified situations as controllable, thus testable, stand-ins for the blooming buzzing confusion of real life. With this truism acknowledged, a symmetrical truism is that the simplifications can, covertly or not, introduce distortions. For example, our current experiments, in company with many other laboratory studies that purport to expose the core problem of word learning, are at best studies of paired association between a speech sound and a small closed set of carefully cropped images devoid of their typical context – i.e., their rich, three-dimensional, moving, interactive, environmental context. We do not know whether, or to what degree, items so acquired generalize across the significant cognitive categories that are usually and necessarily triggered by these more complex contexts. In the real world, we must learn that /dɒg/ means 'dog' even though every observed dog is a different one, seen more or less darkly, barking more or less loudly, and so forth. As is just as well known, and more problematical by far, practically speaking there is no known limit to the number of representations under which any single word-situation pair may be regarded and categorized (cf. Chomsky, 1957; Quine, 1960). As measured in the laboratory, meaning identification is deemed "successful" if the subject has clicked on the correct referent picture, but this engages only part of the problem, as meaning is actually determined by how that referent is intended to be characterized by the speaker.

These considerations suggest that learning a word meaning is unlikely to be a process of statistical elimination of a few competing hypotheses. We can refer to this as the 'open set' problem. Word learning (even, we believe, at its earliest stages) is a mapping problem in which a relatively small set of tokens (words) must systematically connect to a near infinite (open) set of categories. Indeed, it is quite striking that even in such simplified laboratory settings, again and again we see our subjects seizing willy-nilly on a single hypothesis, revising it only if its viability is specifically countermanded by the very next observation. It may very well be that gradualist parallel tracking of associations occurs only within closed or small problem spaces, such as form-to-form mappings within the linguistic system (e.g., syllable co-occurrences, verb subcategorization preferences, etc.). Yet to date, hypotheses have not been adequately tested regarding how these associations are learned in the first place (Roediger & Arnold, 2012).

6.3. Limitations and further questions

There are of course several important issues left unresolved about our propose-but-verify account of word learning. Perhaps the most pressing is whether such a cross-situational learning procedure is also present in infants and young children. There is good reason to suspect that it is, given the striking continuity observed thus far in word learning abilities across children and adults (Gillette et al., 1999), and the cross-species and cross-task observations of one-trial learning just discussed. It is true that in the present word learning procedure (and the procedures used by most others), adult participants are told explicitly how to approach the problem – e.g., they are told that the spoken nonsense words label objects on the screen. Infant participants do not have this benefit, but crucially when linguistic information is provided that such words are referential (e.g., syntactic evidence), infants readily treat the words as referential and face the same challenges associated with referential ambiguity (Fennell & Waxman, 2010; Fulkerson & Waxman, 2007; Namy & Waxman, 2000). Moreover, to the extent that it has been studied, children and adults behave quite similarly in the Human Simulation Paradigm (Medina et al., 2011; Piccin & Waxman, 2007) and even in the artificial referent worlds more typically used in cross-situational word learning studies (L. Smith & Yu, 2008; Yu & L. Smith, 2011).

It is also important to consider further the conditions under which language learners do indeed track multiple hypotheses simultaneously for a word. We suspect, and the current evidence suggests (K. Smith et al., 2011), that multiple-hypothesis tracking occurs under conditions of massed learning, i.e., a series of immediately adjacent encounters with a word in different referential contexts; evidence is also presented in Gillette et al., 1999, who in their first HSP experiments presented all exemplars for each word, rather than (as in later studies) distributed across the full learning set. Immediately-repetitive uses of words do of course sometimes occur in natural speech, e.g. in moments of instruction ("There's a bird! There's another bird!"), and in the so-called "repetition sequences" observed in parents of very young children ("Go get the bird, that's right, the bird, yes, pick up the bird. . ."). Further

work will be needed to see (a) how such repetitions operate in natural child-directed speech and (b) the extent to which multiple-hypothesis tracking under these conditions, if observed, is a reflection of a statistical-associative learning mechanism or simply short-term storage in working memory.

We also suspect that when a confirmed (and even re-confirmed) hypothesis for a word is then not supported by a later context, the learner would actively search memory for past rejected hypotheses, and may even establish a second meaning for the word. Such a mechanism would be needed for the learning of homophones, for example. The establishment of a second meaning would most likely occur when the referential context of a later learning instance does not support a confirmed word meaning and instead unambiguously supports a different meaning (see Vouloumanos, 2008).

Finally, we do not wish to leave the reader with the impression that *propose-but-verify* is a complete model of word learning. Rather it reveals the early stages of the meaning discovery process. The learning of most other vocabulary items (i.e., words that are not concrete count nouns) requires the use of a much broader database of linguistic and nonlinguistic evidence for their accurate discovery. Although one-trial learning may underpin the word-to-meaning discovery for these words as well, such a procedure does not speak to the issue of how information is integrated across linguistic and nonlinguistic domains.

6.4. Closing comments

In partial response to the difficulty of understanding the word learning process, given the twin problems of gaining stimulus control (that is, of devising some situation simple enough to measure at all) and establishing that the situation has some plausible link with reality (that is, addressing the many-to-many mapping problem), we have tried to make some progress by varying these factors across three experiments. In a companion piece to the present one (Medina et al., 2011), we examined cross-situational learning with naturalistic stimuli (40-s muted videos of parent-to-infant speech, in context), thus maintaining some of the real complexity of the child's learning environment. In that situation, participants seemed almost totally unable to remember anything about past learning situations, employing instead the propose-and-verify procedure that requires no tracking of past contexts. The present studies, by reducing this complexity, allowed us to see if something resembling word learning will become cumulative and statistical-associative if reduced to new labeling of image to nonsense word pairings. As we showed, even in this situation, and even though the participant-observers are cognitively mature adults who are familiar with the lexical categories we used, they propose, but verify.

Acknowledgments

This work was funded by the Eunice Kennedy Shriver National Institute of Child Health and Human Development Grant 1-R01-HD-37507 awarded to L.R.G. and J.C.T. We would like to thank members of the Trueswell–Gleitman Language Learning and Language Processing Lab and especially Mariah Moore who assisted us in part of this research.

References

Anglin, J. (1993). Vocabulary development: A morphological analysis. *Monographs of the Society for Research in Child Development, 238*, 1–166.

Arunachalam, S., & Waxman, S. R. (2010). Meaning from syntax: Evidence from 2-year-olds. *Cognition, 114*, 442–446.

Aslin, R. N., & Newport, E. L. (2008). What statistical learning can and can't tell us about language acquisition. In J. Colombo, P. McCardle, & L. Freund (Eds.), *Infant pathways to language: Methods, models, and research directions*. Mahwah, NJ: Lawrence Erlbaum Associates.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*, 390–412.

Baldwin, D. A. (1991). Infants' contribution to the achievement of joint reference. *Child Development, 62*, 875–890.

Baldwin, D. A. (1993). Early referential understanding: Infants' ability to recognize referential acts for what they are. *Developmental Psychology, 29*(5), 832–843.

Barr, D. J. (2008). Analyzing 'visual world' eyetracking data using multilevel logistic regression. *Journal of Memory and Language, 59*, 457–474.

Bates, E., Dale, P. S., & Thal, D. (1995). Individual differences and their implications for theories of language development. In P. Fletcher & B. MacWhinney (Eds.), *Handbook of child language* (pp. 96–151). Oxford: Basil Blackwell.
Bergelson, E., & Swingley, D. (2012). At 6 to 9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences USA, 109*, 3253–3258.
Bloom, P. (2002). Mindreading, communication, and the learning of the names for things. *Mind and Language, 17*, 37–54.
Booth, A. E., & Waxman, S. R. (2002). Word learning is "smart": Evidence that conceptual information affects preschoolers' extension of novel words. *Cognition, 84*, B11–B22.
Bower, G. H. (1961). Application of a model to paired-associate learning. *Psychometrika, 26*, 255–280.
Carey, S. (1978). The child as word learner. In J. Bresnan, G. Miller, & M. Halle (Eds.), *Linguistic theory and psychological reality* (pp. 264–293). Cambridge, MA: MIT Press.
Carey, S., & Bartlett, E. (1978). Acquiring a single new word. *Proceedings of the Stanford Child Language Conference, 15*, 17–29.
Chomsky, N. (1957). *Syntactic structures*. The Hague/Paris: Mouton.
Duffy, S. A., Morris, R. K., & Rayner, K. (1988). Lexical ambiguity and fixation times in reading. *Journal of Memory and Language, 27*, 429–446.
Estes, W. K. (1956). The problem of inference from curves based on grouped data. *Psychological Bulletin, 53*, 134–140.
Estes, W. K. (1960). Learning theory and the new "mental chemistry". *Psychological Review, 67*, 207–223.
Estes, W. K. (2002). Traps in the route to models of memory and decision. *Psychonomic Bulletin & Review, 9*, 3–25.
Fennell, C., & Waxman, S. R. (2010). What paradox? Referential cues allow for infant use of phonetic detail in word learning. *Child Development, 81*(5), 1376–1383.
Fisher, C., Gertner, Y., Scott, R., & Yuan, S. (2008). Verb learning and the early development of sentence comprehension. *International Journal of Psychology, 43*(3–4), 177.
Fisher, C., Hall, S., Rakowitz, L., & 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. A. (1983). *Modularity of mind: An essay on faculty psychology*. Cambridge, Mass.: MIT Press.
Frank, M. C., Goodman, N. D., & Tenenbaum, J. B. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science, 20*, 579–585.
Frank, M. C., Tenenbaum, J. B., & Fernald, A. (in press). Social and discourse contributions to the determination of reference in cross-situational word learning. *Language Learning and Development*.
Fulkerson, A. L., & Waxman, S. R. (2007). Words (but not tones) facilitate object categorization: Evidence from 6- and 12-month-olds. *Cognition, 105*, 218–228.
Gallistel, C. R., Balsam, P. D., & Fairhurst, S. (2004). The learning curve: Implications of a quantitative analysis. *Proceedings of the National Academy of Sciences USA, 101*(36), 13124–13131.
Gebhart, A. L., Newport, E. L., & Aslin, R. N. (2009). Statistical learning of adjacent and non-adjacent dependencies among non-linguistic sounds. *Psychonomic Bulletin & Review, 16*, 486–490.
Gervain, J., Nespor, M., Mazuka, R., Horie, R., & Mehler, J. (2008). Bootstrapping word order in prelexical infants: A Japanese–Italian cross-linguistic study. *Cognitive Psychology, 57*, 56–74.
Gillette, J., Gleitman, H., Gleitman, L. R., & Lederer, A. (1999). Human simulations of vocabulary learning. *Cognition, 73*, 135–176.
Gleitman, L. R. (1990). Structural sources of verb learning. *Language Acquisition, 1*, 1–63.
Gleitman, L. R., Cassidy, K., Nappa, R., Papafragou, A., & Trueswell, J. C. (2005). Hard words. *Language Learning and Development, 1*(1), 23–64.
Gómez, R. L., & Gerken, L. A. (2000). Infant artificial language learning and language acquisition. *Trends in Cognitive Sciences, 4*, 178–186.
Gropen, J., Pinker, S., Hollander, M., & Goldberg, R. (1991). Affectedness and direct objects: The role of lexical semantics in the acquisition of verb argument structure. *Cognition, 4*, 153–195.
Heibeck, T. H., & Markman, E. M. (1987). Word learning in children: An examination of fast mapping. *Child Development, 58*, 1021–1034.
Ichinco, D., Frank, M. C., & Saxe, R. (2009). Cross-situational word learning respects mutual exclusivity. In *Proceedings of the 31st annual meeting of the cognitive science society*.
Jaswal, V. K. (2010). Explaining the disambiguation effect: Don't exclude mutual exclusivity. *Journal of Child Language, 37*, 95–113.
Kachergis, G., Yu, C., & Shiffrin, R. (2010). Adaptive constraints and inference in cross-situational word learning. In *Proceedings of the 32nd annual conference of the cognitive science society*.
Klein, K., & Yu, C. (2009). Joint or conditional probability in statistical word learning: Why decide? In *Proceedings of the 31st annual conference of the cognitive science society*.
Landau, B., & Gleitman, L. R. (1985). *Language and experience: Evidence from the blind child*. Cambridge, MA: Harvard University Press.
Landau, B., Smith, L. B., & Jones, S. S. (1988). The importance of shape in early lexical learning. *Cognitive Development, 3*, 299–321.
Lashley, K. S. (1942). An examination of the "continuity theory" as applied to discriminative learning. *The Journal of General Psychology, 26*, 241–265.
Maratsos, M., & Chalkley, M. (1981). The internal language of children's syntax. In K. E. Nelson (Ed.). *Children's language* (vol. 2). New York: Gardner Press.
Markman, E. M. (1989). *Categorization and naming in children: Problems of induction*. MIT Press.
Markson, L., & Bloom, P. (1997). Evidence against a dedicated system for word learning in children. *Nature, 385*, 813–815.
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 USA, 108*, 9014–9019.
Miller, G. A. (1963). Comments on professor postman's paper. In C. N. Cofer & B. S. Mugrave (Eds.), *Verbal behavior and learning: Problems and processes* (pp. 321–329). New York, NY: McGraw-Hill Book Company.
Namy, L. L., & Waxman, S. R. (2000). Naming and exclaiming: Infants' sensitivity to naming contexts. *Journal of Cognition and Development, 1*, 405–428.

Nappa, R., Wessell, A., McEldoon, K. L., Gleitman, L. R., & Trueswell, J. C. (2009). Use of speaker's gaze and syntax in verb learning. *Language Learning and Development, 5*(4), 203–234.
Naigles, L. (1990). Children use syntax to learn verb meanings. *Journal of Child Language, 17*, 357–374.
Newport, E. L., & Aslin, R. N. (2004). Learning at a distance: I. Statistical learning of non-adjacent dependencies. *Cognitive Psychology, 48*, 127–162.
Osgood, C. E., Suci, G., & Tannenbaum, P. (1957). *The measurement of meaning*. Urbana, IL: University of Illinois Press.
Papafragou, A., Cassidy, K., & Gleitman, L. R. (2007). When we think about thinking: The acquisition of belief verbs. *Cognition, 105*(1), 125–165.
Piccin, T. B., & Waxman, S. R. (2007). Why nouns trump verbs in word learning: New evidence from children and adults in the Human Simulation Paradigm. *Language Learning and Development, 3*(4), 295–323.
Quine, W. V. (1960). *Word and object*. The MIT Press.
Rock, I. (1957). The role of repetition in associative learning. *American Journal of Psychology, 70*, 186–193.
Roediger, H. L., III, & Arnold, K. M. (2012). The one-trial learning controversy and its aftermath: Remembering Rock (1957). *American Journal of Psychology, 125*(2), 127–143.
Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month old infants. *Science, 274*, 1926–1928.
Simpson, G., & Burgess, C. (1985). Activation and selection processes in the recognition of ambiguous words. *Journal of Experimental Psychology: Human Perception and Performance, 11*(1), 28–39.
Siskind, J. M. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition, 61*(1–2), 1–38.
Smith, K., Smith, A. D. M., & Blythe, R. A. (2011). Cross-situational learning: An experimental study of word-learning mechanisms. *Cognitive Science, 35*, 480–498.
Smith, L. B., Yu, C., & Pereira, A. F. (2009). Not your mother's view: The dynamics of toddler visual experience. *Developmental Science* 1–9.
Snedeker, J. (2009). Word learning. In L. R. Squire (Ed.), *Encyclopedia of neuroscience* (pp. 503–508). Amsterdam: Elsevier.
Snedeker, J., & Gleitman, L. (2004). Why it is hard to label our concepts. Hall & Waxman (Eds.), *Weaving a lexicon*. Cambridge, MA: MIT Press.
Soja, N., Carey, S., & Spelke, E. S. (1991). Ontological categories guide young children's inductions of word meaning: Object terms and substance terms. *Cognition, 38*, 179–211.
Steiger, J. H., Shapiro, A., & Browne, M. W. (1985). On the multivariate asymptotic distribution of sequential chi-square statistics. *Psychometrika, 50*, 253–264.
Stevens, J., Yang, C., Trueswell, J. C., & Gleitman, L. R. (2012). Learning words via single-meaning hypothesis testing. *Paper presented at the 86th Annual Meeting of the Linguistics Society of America*. Portland, OR.
Trueswell, J. C. (1996). The role of lexical frequency in syntactic ambiguity resolution. *Journal of Memory and Language, 35*, 566–585.
Trueswell, J. C., & Gleitman, L. R. (2007). Learning to parse and its implications for language acquisition. In G. Gaskell (Ed.), *Oxford handbook of psycholinguistics* (pp. 635–656). Oxford, England: Oxford University Press.
Vouloumanos, A. (2008). Fine-grained sensitivity to statistical information in adult word learning. *Cognition, 107*, 729–742.
Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review, 114*(2), 1087–1094.
Yu, C. (2008). A statistical associative account of vocabulary growth in early word learning. *Language Learning and Development, 4*(1), 32–62.
Yu, C., & Smith, L. B. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science, 18*(5), 414–420.
Yu, C., & Smith, L. B. (2011). What you learn is what you see: Using eye movements to study infant cross-situational word learning. *Developmental Science, 14*(2), 165–180.
Yu, C., & Smith, L. B. (2012). Modeling cross-situational word-referent learning: Prior questions. *Psychological Review, 119*(1), 21–39.
Yuan, S., & Fisher, C. (2009). "Really? She blicked the baby?": Two-year-olds learn combinatorial facts about verbs by listening. *Psychological Science, 20*, 619–626.