HHS Public Access

Author manuscript

Top Cogn Sci. Author manuscript; available in PMC 2021 January 01.

Published in final edited form as:

Top Cogn Sci. 2020 January ; 12(1): 22–47. doi:10.1111/tops.12352.

Easy Words: Reference Resolution in a Malevolent Referent World

Lila R. Gleitman, **John C. Trueswell**

Department of Psychology, University of Pennsylvania

Abstract

This article describes early stages in the acquisition of a first vocabulary by infants and young children. It distinguishes two major stages, the first of which operates by a stand-alone word-to-world pairing procedure and the second of which, using the evidence so acquired, builds a domain-specific syntax-sensitive structure-to-world pairing procedure. As we show, the first stage of learning is slow, restricted in character, and to some extent errorful, whereas the second procedure is determinative, rapid, and essentially errorless. Our central claim here is that the early, referentially-based learning procedure succeeds at all because it is reined in by attention focusing properties of word-to-world timing and related indicants of referential intent.

Keywords

psycholinguistics; language development; word learning; vocabulary growth; reference

1. Introduction

In 2005, our research group outlined an updated theory of how language learners acquire what we called “hard words” (Gleitman, Cassidy, Nappa, Papafragou, & Trueswell, 2005). “Hard” and “easy” here pertain to the facts about acquisition. “Acquisition” in the regards we will discuss means gaining the knowledge that the concept ‘dog’ is expressed by the sound segment /dɒg/ in English[1]. As this learning process begins, it seems reasonable to suppose that the child’s only recourse for solving this problem is to examine the relation between some recurrent sound segment /dɒg/ and what is happening out in the world (in the best case, dog sightings). Hard words, roughly, are the ones whose meanings do not come readily to mind as a consequence of this word-to-world pairing procedure. On this definition the word *think* would be classified as “hard” because observers do not usually guess “They’re thinking” upon observing thinkers thinking. In contrast, the word *jump* would be classified as “easy” because observers, quite consensually, guess “They’re jumping” upon observing jumpers jumping. And indeed language learners acquire words like *dog* and *jump* before words like *idea* and *think*. We proposed that the acquisition of the latter require the use of linguistic context, which must be organized as a structured syntactic representation; it

Send Correspondence to: Lila Gleitman and John Trueswell, Department of Psychology, University of Pennsylvania, Steven A. Levin Building, 425 S University Avenue, Philadelphia, PA 19104, gleitman@psych.upenn.edu, trueswel@psych.upenn.edu.

[1]Notationally, we use italics for mention of a phrase or word, “double quotes” for its utterance or sound, and ‘single quotes’ for a concept.

is the easy words that allow the learner to construct those representations. Putting this theory in shorthand terms, the postulate was that learners could align their sampled structured representations (syntactically analyzed sentences of, say, English) with the counterparts of these in a structured representation of the observed situation - a procedure called "syntactic bootstrapping" (Gleitman, 1990; Landau & Gleitman, 1985). This approach found support and was materially fleshed out in a series of empirical demonstrations that followed over the last few decades (e.g., Nappa, Wessell, McEldoon, Gleitman & Trueswell, 2009; Papfragou, Massey, & Gleitman, 2006; Snedeker & Gleitman, 2003; Yuan, Fisher, & Snedeker, 2012) and, indeed, had been documented in many of its elements by several demonstrations and theoretical commentaries that led up to it (e.g., Fisher, Gleitman & Gleitman, 1991; Fisher 1996; Gleitman, 1990; Gleitman & Landau, 1994; Lidz, Gleitman, & Gleitman, 2003; Naigles 1990, 1996; Naigles, Gleitman & Gleitman, 1986; for reviews, Gleitman & Fisher, 2005; Fisher & Gleitman, 2002).

In retrospect, progress on the hard-word front might have been expected because the explanation drew upon, and was couched in, the kind of theoretical apparatus and terminology that Cognitive Science was designed for and proved good at (cf., Turing, 1950; Chomsky, 1959; Fodor, 1975; Pinker, 1984; Wexler & Collicover, 1980). So, for instance, Fisher, Gleitman, and Gleitman (1991) could predict with relative confidence why *laugh* (which describes a self-caused action) occurred in one-argument (intransitive) sentences, and by so doing recover the argument-taking properties of this predicate in a formally satisfying way, with obvious pointers to a learning schema. But these same authors acknowledged that this approach casts little light on the "other part" of the meaning of *laugh*, namely, the "ha-ha" part. Contemporary critics, and the authors of such theories themselves, looked hopefully to "the world of observation" (how words line up with their extralinguistic contingencies) to solve this "other part" of the lexical acquisition problem, sometimes being unkind enough to remark that the "ha-ha" part was what, after all, everybody else had taken to *be* the meaning of *laugh* in the first place (cf., Pinker, 1994; Fodor, personal communication) and that we are light-years from a theory of this part. Perhaps paradoxically, the easy words and the easy parts of the hard words are the more resistant to current theorizing and explanation.

In the present essay, we review some of the contemporary thinking and findings on this problem: how the "easy words" (and the "easy" parts of hard words) could be identified by a suitably endowed organism - by hypothesis, the human infant who is attempting to map words to their referents. But first, the problem and why it is so hard.

1.1 The hard problem of learning easy words from observation

We have just asserted (as if it were a truism) that in order for language learners to acquire the meanings of their first words, they must rely exclusively on what they can observe from the immediate situational context. At this initial stage, by definition, they don't have internal linguistic information to help with the task. For instance, from an utterance like "The chef baked the cake" it would be easier to learn the meaning of the word *bake* if you had previously learned the meaning of *chef* and *cake*. And it would be easier still if you knew that "chef" was the Subject of the sentence, in which case you might conjecture the meaning

'bake', whereas if you knew "cake" was the Subject you might conjecture 'poisoned'. Instead, learners without such knowledge must see (hear, feel, smell) things that exist out in the world, hear a speech segment, and hopefully make the correct connection. Speaking practically, how hard is this? Figure 1, which shows the vocabulary growth of the child, provides some hints to the answer to this question. Infant word learning appears to be slow and laborious until the dawn of syntax. For the first 100 words or so, which are likely acquired from observation alone (Lenneberg, 1967; Gleitman et al., 2005), the rate of vocabulary accrual is unimpressive at best. Even after 18 months, the child's vocabulary size is still only about 100 words. But a dramatic inflection point is observed immediately thereafter, when the first rudiments of connected speech (and by implication syntax) are observed. Very similar patterns are observed in estimates of child receptive vocabulary (see Caselli, Bates, Casadio, Fenson, Fenson, Sanderl & Weir, 1995). And indeed, as many have noted before (e.g., Chomsky, 1959; Quine, 1960), learning one's first words is not an easy problem to solve even if we make reasonable, evidentially supported assumptions.

The primary critical assumption is that even early word learning is a mapping problem between two kinds of categories (linguistic categories on the one hand, and conceptual categories on the other), and not, for example, an unprepared multi-modal sensory associative process. The child who is prepared to learn words must, for example, already have candidate word forms to work from - i.e., linguistic categories comprised of syllables and phonemes that permit correct generalization across speakers and permit easy detection within the continuously varying speech stream. This preparation presumably takes about 6–8 months of exposure to the language to start a candidate lexicon of word forms (e.g., Johnson & Jusczyk, 2001; Thiessen & Saffran, 2003; Saffran, Aslin & Newport, 1996) and morphosyntax (Marcus, Vijayan, Bandi Rao, & Vishton, 1999). On the world side of the word-to-world equation, we also cannot assume that a learner's mental representation is an unprepared sensory stimulation: just like speech sounds differ from different speakers on different occasions, a dog sighting stimulates the mind in different ways every time it is experienced — different angle, different lighting, even a different dog. At the time of word learning, we have assume these different circumstances all evoke the category 'dog'.[2]

Here though is the where the problem of early word learning gets hard. Very hard. Consider the situation illustrated in Figure 2, in which a boy is hearing the word *shoe* uttered in a typical context. If this confrontation with English is to be of use to this word learner, he must confront the problem of reference resolution, the main burden of the present article. As Figure 2 illustrates, shoes do not usually appear on white backgrounds in the form of a still photograph; there are many other objects, events, and qualia present with any particular shoe sighting. If the learner duly records every entity, relation, or implied motion in this scene, the

[2]There is great disagreement as to what could be meant by the concept 'concept' in the first place. We cannot of course engage this enormous and ill-formulated problem here. Suffice to say, some have thought that concepts are mental images (Hume). Others have thought that a concept is a definition (a set necessary and sufficient conditions for the object to fall under this concept, Katz & Fodor, 1963). And still others have thought it is the typical or prototypical usage conditions (Rosch, 1975; Rosch & Mervis, 1975). For present purposes, we can remain blissfully neutral in these controversies, but granting on the principle of charity that the child's success depends on reaching consensus with the adult world on such representations. Some commentators hold that child representations differ somewhat or perhaps even radically from adults and change as a function of concept growth and acquisition (Carey, 1978; Gentner, 1982; Smiley & Huttenlocher, 1995). Whereas others hold that information availability rather than concept change accounts for the changing character of early vocabulary (Gillette, Gleitman, Gleitman & Lederer, 1999).

task of learning looks hopeless. But a first clue is that the child is looking toward the shoes. Perhaps we can understand the problem of reference resolution by concentrating on such social-attentive cues. Several researchers have provided positive evidence in this regard. Bruner (1974/1975) usefully pointed out that speaker and child are in cahoots, concentrating their attention jointly on only certain elements in the scene. This suggestion has been backed up by a wide range of experimental findings demonstrating that the presence of these social-attentional cues to reference facilitate infant and child word learning (e.g., Baldwin, 1991, 1993, 1995; Bloom, 2000; Jaswal, 2010; Nappa, Wessell, McEldoon, Gleitman, & Trueswell, 2009; Tomasello & Farrar, 1986; Southgate, Chevallier, & Csibra, 2010; Woodward, 2003; for a recent review, see Papafragou, Friedberg, & Cohen, 2017). In her classic study, Baldwin (1991) found that 16–19 month old infants are sensitive to a parent's attentional stance (cued via eye gaze, head- and body-posture) when identifying the referent and learning the meaning of a novel word. Infants connected the parent's utterance of the word to an attended object if and only if the parent was also attending to that object.

An optimistic picture that could emerge from this body of evidence is that early word learning, so filtered, is not so hard after all. If social attentional alignment between the parent and child is commonplace - the rule rather than the exception - it is possible that a large proportion of word occurrences are informative learning moments. Learners who are aggregating across these instances would be able to identify reference and meaning rapidly as they get convergence (Yu & Smith, 2007). And yet, until quite recently (e.g, Clerkin, Hart, Rehg, Yu & Smith, 2017; Roy, Frank & Roy, 2009), surprisingly little research has asked what the situational context is like in the home for children learning their first words. This is where our inquiries began (see Section 2: What is the situational context like in the home?*)*. It is likely that knowing what the informativity patterns are in the home will allow us to offer better theories of the cognitive machinery that supports word learning (a topic we will take up in Section 3: What kind of learner operates on this input?).

2. What is the situational context like in the home?

> For if we will observe how children learn languages, we shall find that, to make them understand what the names of simple ideas or substances stand for, people ordinarily show them the thing whereof they would have them have the idea; and then repeat to them the name that stands for it; as white, sweet, milk, sugar, cat, dog.
>
> (John Locke, 1690, Book 3, IX.9)

The laboratory work on reference resolution mentioned in the previous section establishes that young language learners can use several social-attentional cues to determine the referential intent of a speaker and thus acquire an unfamiliar word's meaning. Moreover, observational research examining parent-child interactions has established that parents who spontaneously exhibit these behaviors during object play, e.g., of labeling what a child is attending, have children with larger vocabularies (e.g., Harris, Jones, Brookes, & Grant, 1986; Tomasello & Farrar, 1986; Tomasello, Mannle, & Kruger, 1986; Tomasello & Todd, 1983). However, neither the laboratory nor the observational studies indicate what the rate of informative learning instances is in the home under more common circumstances. The

observational work has been almost exclusively limited to object play, typically with a small set of predetermined objects. Moreover, with few exceptions, both the laboratory work and the observational work has been limited to situations when parents utter nouns with concrete meanings. This certainly overestimates the informativity of word learning situations. A child learning her first words does not know which ones have concrete meanings and which do not, nor are they likely to know which label objects and which do not. If we expand our assessment of referential informativity to content words in general (i.e., recurring word forms that receive some kind of prosodic stress above and beyond function words), and we expand our assessment to more than just object play, it is entirely possible that these informative situations are exceedingly rare and atypical.

2.1 Rarity of referential gems and their contribution to vocabulary growth

Our own work suggests that moments of referential clarity are very rare in the home. Much of the evidence comes from a procedure we have called the Human Simulation Paradigm (HSP) first introduced in Gillette, Gleitman, Gleitman, and Lederer (1999). The authors suggested that different sources of evidence as to an unfamiliar word's meaning might be available from various aspects of the input stream. In principle, as we have discussed so far, the situational context ("observation") would supply some evidence; later, the co-occurrence of the unfamiliar word with familiar ones ("distribution") could supply further evidence; and later still, the structural position of the unfamiliar word in the sentence ("syntax") could provide yet even more evidence. An experimental procedure was developed to estimate the relative contribution of these several sources of evidence taken alone or jointly. Specifically, a video corpus was built of examples of parents uttering common nouns and verbs within natural sentences to their offspring, partitioned into 40-second "vignettes" in which the word of interest (henceforth, the "mystery word") had been uttered by the adult 30 seconds into each vignette. For each mystery word, six vignettes were chosen randomly from the corpus. One group of participants observed the vignettes with the sound off with the mystery word indicated by an audible beep at its exact occurrence, a proxy for the extra-linguistic situation taken alone ("observation"). A second group saw just "distributional" evidence for the mystery word, i.e., an alphabetical list of the nouns as they occurred in each of these six sentences (e.g., "BLEPPED: cake, chef" for the sentence "The chef BAKED the cake."). A third group was provided with "jabberwocky" versions of the parent's utterances, i.e., with content words replaced with nonsense words ("syntax": "The florp BLEPPED the dax"). Further participant groups received two or all three of these kinds of information together (see especially Snedeker & Gleitman, 2003). In all of these conditions, participants were given six examples of any one word in a row and, in Gillette et al., participants were told in advance whether the mystery word was a noun or a verb. The central finding of this work was that the different participant groups learned different aspects of vocabulary based on these different input representations. Those who received only the observational evidence learned mainly concrete nouns and not verbs, thus reproducing in adults the first stages of vocabulary acquisition by infants (e.g., as documented in Bates, Dale, & Thal, 1995). Adults who were provided with the "distribution" and especially "syntax" information successfully identified more abstract words including the verbs and particularly those with abstract meaning, thus reproducing the aspects of later child vocabulary. The bottom line conclusion was that the course of early vocabulary learning is legislated by the order in which the child

has access to the different sources of information and not perhaps by some hypothesized conceptual change in the nature of the learner - information change rather than conceptual change. The child hears sentences in situational contexts from the beginning but appreciates the distributional and syntactic cues only at later stages.

Perhaps the most relevant aspect of the Gillette et al. (1999) findings to the present discussion is that the initial learning procedure ("observation") was somewhat unimpressive even in its own terms. Adults in the "observation" condition, who were given a video recording of the local, seemingly quite relevant, situational context, were only able to guess the correct meaning of 40% of the most common nouns and 15% of the most common verbs. Even this is an overestimate of the informativity of the observational database because: 1) all six vignettes of, e.g., 'dog'-mentionings, were presented in sequence, after which learners offered a final guess (this massed-trial procedure is surely not a feature of everyday conversation); and 2) participants were actually told whether the mystery word was a noun or a verb (again surely not information explicitly provided to real learners). This raises the question of how informative observational evidence actually is for first words.

To this end, our group set out to understand better the true patterning of parental word use in the home that might be underpinning the acquisition of first words. Medina, Snedeker, Trueswell and Gleitman (2011, Exp 1) used the "observational" condition of the HSP to answer this question. Medina et al. improved the procedure of Gillette et al. (1999). They used a new vignette corpus that sampled more widely across common daily activities, such as bath, meal and play time. As in Gillette et al., the resulting video corpus consisted of 288 vignettes in which parents uttered common content nouns or common content verbs (144 vignettes each, 24 word types each, 6 vignettes for each word). Videos were again muted and a beep occurred just as the parent had uttered the mystery word. Unlike as in Gillette et al., participants were not told which words were nouns and which were verbs, nor were the six examples of a word strung together in a row; vignettes were intermixed and appeared in a random order, eliminating the benefit of comparing information from successive instances of a single word. Under these more natural sampling conditions, nouns were guessed correctly only 17% of the time and verbs a paltry 6% of the time. Medina et al. also noted that not all word exposures (i.e., vignettes) are created equal. Only a small percentage (7%) were informative above the 50% accuracy rate. We can think of these as referential "gems", in which the situational context, including the social interaction between parent and child offered enough information to allow for the majority of naive observers to guess what the parent was saying. Strikingly, all gems were nouns, and none verbs. Moreover, most other vignettes might be characterized as referential "junk", yielding a less than 33% accuracy rate. Worse, observers typically offer a scatter of different false conjectures rather than one or two (as also observed in Gillette et al.).

The picture that emerges, then, is a learning environment in which most instances of a given word's use are uninformative but punctuated by rare occasions of referential clarity. But before taking these findings and conclusions at face value, a pressing question that must be answered is whether adult responses in the HSP are in fact an adequate proxy for the referential experiences of the children in these videos. After all, the videos are taken from a

third-person angle and judgments were from adults rather than children. At least three findings in the literature alleviate these concerns:

(1) Children perform similarly to adults on the HSP task.—It is possible that responses from adults in the HSP over-estimate the complexity of referential contexts, since adults may possess more concepts and more ways of interpreting scenes. Children may consider far less information and thus may be better at identifying referents if they preferentially select common word meanings —a form of the "less is more" argument (Elman, 1993; Newport, 1990). However, results do not support this possibility. Adult HSP viewers tend to guess meanings corresponding to children's first words (Gillette et al., 1999). Moreover, child-friendly HSP studies have been conducted, in which children (4 years of age and older) produce results similar to adults (Medina et al., 2011; Piccin & Waxman, 2007). Rather than doing better than adults, children perform worse overall (Medina et al.) but nevertheless present similar patterns in their data. In particular, similar to adults, they are more accurate on nouns than verbs (Piccin &Waxman) and find the same contexts highly informative (Medina et al.).

(2) Third person angle is informative.—The child's own view in these HSP vignettes is quite different from the one offered to HSP observers who viewed the situation from a 3rd person camera angle. If HSP observers were not focusing on where the child was looking in these videos, it is possible that HSP observers consider very different information than the children in these videos. After all, the 3rd person videos step back and show a rather broad view of the whole scene, while the child's focused vision is much more restrictive (Smith, Colunga, & Yoshida, 2010; Smith, Yu & Pereira, 2011). The question of course is whether that distinction, dramatic as it seems, makes a difference. Yet, when compared, HSP responses from videos from a 1st-person head-mounted child camera yield very similar accuracy results to responses from videos of the same parent-child interactions viewed from a 3rd person fixed camera angle (Yurovsky, Smith & Yu, 2013). In particular, HSP accuracy was exactly 58% correct from both angles and yielded very similar distributions of accuracy across items. (Higher accuracy was observed here than all other HSP studies presumably because only object play was considered and only utterances that referred to co-present objects were selected as stimuli.) First person advantages were very small and were found only in a second study using the least informative learning instances. The lack of substantial differences between 1st and 3rd person camera angles is less surprising when one considers that adult observers are able to judge what another person is looking at under live-action conditions in which head-turn and gaze information are present (Doherty, Anderson, & Howeieson, 2009).

(3) Effects of input conditions on vocabulary size and growth.—Now we turn to the truly newsworthy outcomes of such studies - that is, evidence of real-world applicability of these laboratory HSP findings (Cartmill, Armstrong, Gleitman, Goldin-Meadow, Medina, & Trueswell, 2013). Cartmill et al. compiled video vignettes of parents uttering common concrete nouns to their 14- to 18-month olds sampled from 56 different families ranging in SES (as part of a larger longitudinal study, see e.g., Goldin-Meadow, Levine, Hedges, Huttenlocher, Raudenbush & Small, 2014). Cartmill et al. were able to show that differential

informativity (that is, the ratio of gems to junk) varies greatly across families and predicts measures of vocabulary size when assessed three years later at school entry. Specifically, for the most referentially transparent families, HSP participants guessed the parent's intended meaning 45% of the time, and for the most referentially opaque families only 5% of the time! This difference predicted Peabody Vocabulary test scores measured at kindergarten-entry, with this relationship holding even after controlling for the amount of talk in the home (see Figure 3). Cartmill et al. found that both quantity of talk and its quality (i.e., its referential transparency) predict vocabulary growth. Notably, although the amount of talk was found to be positively related to family income (as observed in many studies previously, e.g., Hart & Risley, 1995), the difference in the proportion of gems to junk is a familial and not an SES variable. Although more advantaged families talk more to their children, they do not provide as a group a larger proportion of informative learning instances. What this means commonsensically is that it matters whether you are talking *with* your children rather than *at* them. That is, the difference comes down to commenting on the visibly passing scene more than on beliefs, desires, and commands pertaining to the world at large.

2.2 Characterizing referential gems

Another pressing question that all of this work raises is what exactly makes a referential act a gem? If laboratory evidence is a good guide, one might expect that social-attentive behaviors present in these videos determine moments of referential clarity. Trueswell, Lin, Armstrong, Cartmill, Goldin-Meadow, and Gleitman (2016) reported a large-scale video coding project of the Cartmill et al. (2013) HSP vignettes and confirmed that social-attentive behaviors are indeed associated with referential clarity, but they also identified a crucial amendment to this conclusion: *it is the precise timing of these cues to reference, not their mere presence, that determines referential clarity*. For each 40 second HSP vignette from the Cartmill et al. (2013) study, trained coders annotated the videos on a moment-by-moment basis for the presence of social-attentive cues to reference, including referent presence, parent attention to the referent, child attention to the referent, and parent gesture/manipulation of the referent. The timing of this information was found to predict the HSP accuracy score for each vignette. For instance, although low informative vignettes (junk) were overall less likely to have the referent object present during the interaction as compared to high informative gems, Figure 4 shows that it is really the sudden appearance of the referent just prior to its mention that is an informative cue.

The importance of the timing was also observed for social-attentive cues to reference, such that: (1) gems, not junk, are more likely to contain a sudden shift in attention by parent or child just prior to the word onset (Figure 5A and 5B), which can also be characterized as (2) the onset of joint attention between adult and child (Figure 5C); (3) parent gestures to and manipulation of the referent at or near the time of word onset are also associated with referential gems (Figure 5D). A second-by-second multiple regression that entered all these cues simultaneously to predict HSP accuracy scores confirmed that all of these cues together contribute to HSP accuracy except for gesture / manipulation, which no longer was significant. What this means is that no single cue is a driving factor for referential clarity. For instance, Trueswell et al. found that even if one analyzes only those videos in which the referent was present throughout the entire video, the exact same timing patterns emerge for

the social-attentive cues or, in other words, Figure 5 would look very similar if it included only vignettes in which the referent was present throughout. Such patterns suggest that exact timing of information, not just the continued presence of it, drives referent identification.

Trueswell et al. (2016) asked much more specifically about how these timing relations influence learning. If "good timing" supports learning, it follows that "bad timing" should defeat it. If the relation between the word and world must be exquisitely tight, as in the imputation of cause-and-effect in the physical world (e.g, Michotte, 1963; Leslie & Keeble, 1987), this would affect a massive further filter on the learning procedure. Accordingly, Trueswell et al. (2016), in a smaller HSP study of their own, experimentally manipulated the temporal relation between the observed world and the word form, while keeping the broader situational content constant. The method was simple, gems from previous HSP studies were modified in time such that the "mystery word" was uttered not when the linguistic event actually happened but either 1, 2 4 or 6 seconds before or after. Does this matter? If the general gist of the situation in its entirety were driving the subject's conjecture then one or a few seconds mismatch should not matter much. But if the timing relationship is required to be very tight we should see decrements in confidence and correct guessing in the face of even one or two seconds of mismatch. Indeed, this is what was observed (see Figure 6). A corresponding study with children replicates this pattern in the essential target population (Trueswell, Dawson, Pozzan & Gleitman, in prep.).

2.3 Interim summary

We believe these results offer a better picture of the informational composition and distribution of referential acts in the home for young children learning their first words, more so than studies that are pre-selective both in terms of stimulus items and presentation conditions. Our findings suggests that rare moments of referential clarity exist in the everyday uses of common content words. Most other moments are referentially opaque in the sense that even adult observers of these parent-child interactions are unable to guess what is meant by a speech act when deprived of the linguistic signal. Moments of referential clarity are characterized by precisely timed social-attentive cues to reference. Moreover, homes with a greater proportion of moments of referential clarity have children whose vocabularies outpace children in homes with fewer highly informative interactions. These results point to a possible solution for how learners could solve the indeterminacy problem (so many referents in the world, yet the word refers to only one on a single occasion). If a learner has means for distinguishing the gems from the junk (a topic we take up in the next section), then one or two gems may be enough for referent identification. In other words, massive indeterminacy is a fact about the world but may not matter to the suitably endowed learner.

3. What kind of learner operates on this input?

> Nobody joins the voice of a sheep with the shape of a horse, nor the color of lead with the weight and fixedness of gold…unless he has a mind to fill his head with chimeras, and his discourse with unintelligible words.
>
> (John Locke, 1690, Book 3, VI.28)

We have so far given evidence that the input database for the novice learner taken at face value seems malevolent: even though there are gems, there is an enormous amount of junk. How is the learner to navigate through input of this kind? At one extreme, the learner could accept all of the information gleaned from gems and junk alike, winnowing down the meaning conjectures by locating recurrent sub-patterns in a sea of chaos. At bottom, this is an associative procedure in the sense first outlined by Hume (1748) and accepted by many present day theorists who propose cross-situational models of word learning in which all available data is aggregated (Siskind, 1996; Yu & Smith, 2007; inter alia). What this means however is that a wide range of erroneous information will be entered into the learning system every time "junk" is encountered. For instance, as Locke points out (but does not recommend), you could take the co-occurrence of a passing horse accompanied by a voice of a sheep heard nearby to construct some mythical representation of a sheephorse, or you could say there are two words one approximating the meaning of sheep and the other the meaning of horse. But we are recommending a third way (which was Locke's way too). Impressed by the apparent facts about real learning, we have proposed that the learner enters into the lexicon only a single conjecture on any learning instance, gem or junk. On the succeeding occurrence of this word form, that conjecture is tested for fit to the current situation, in a process we have called propose-but-verify. It is interesting to consider what these two approaches, cross-situational vs. propose-but-verify, are expected to do on a moment of referential junk. The former will enter into learning a wide variety of erroneous conjectures (e.g., *sheep, horse, sheephorse)* whereas the latter will store just one, thereby limiting the situation's negative impact. In propose-but-verify, a checking procedure does the rest; the next time the word is encountered the learner retrieves the original conjecture (be it *horse, sheep or sheephorse)* and attempts to verify it against the new stimulus. This procedure is effective for it is unlikely in the extreme that one will again hear the voice of a sheep upon seeing the shape a horse.

3.1 Targeted learning under cross-situational circumstances

The first evidence for this highly targeted learning procedure came from an HSP study that permitted the subjects to engage, if they wished, in cross-situational comparison (Medina et al., 2011, Exp. 2). The stimuli were vignettes of parents uttering common nouns to their young children in the home under everyday circumstances. Five vignettes were selected for each word. Instead of hearing a "beep" at the exact moment the target word was uttered, adult observers heard a nonsense word (e.g., for each "ball" vignette participants heard "mipen", for each "horse" vignette, "dax"). For each vignette, participants wrote down their best guess as to the meaning of the nonsense word, but here the recurring instances of a nonsense word were opportunities for cross-situational comparison, assuming participants could recall aspects of the previous "mipen" vignette(s). We distributed referential clarity across word occurrences (the distribution of gems and junk) to approximate what is found in the natural input: only one in five instances of each nonsense word was a gem, and sometimes no gem occurred at all. As in natural input, we intermixed the word types such that not all "mipen" vignettes were encountered in a row.

Participants' successive guesses for each word suggested that they were following a learning procedure like propose-but-verify. In propose-but-verify, the learner is expected to pair a

word form with a single referent (the current proposal) but not connect the word with any other information about that situation, including other hypothesized referents. So while the learner might very well remember that he saw the horse in a corral on an early Sunday morning with a ribbon around its neck, and even as a sheep passed by, these aspects of the situation are not stored alongside the word for the purposes of the acquisition of the emerging lexicon. Upon encountering the word "horse" for a second time, the learner retrieves from memory what they think "horse" means, not the past situation in which the word "horse" was encountered. We see this in our own data (Figure 7). When people made an incorrect guess as to the referent to the word, their performance on the next word's occurrence was no better than other participants who had viewed the same vignette in isolation without the benefit of cross-situational comparison. If participants had recalled the past situation, they would have been expected to outperform those subjects who saw only a single instance, but they did not. For instance, when the mystery word was "horse", horses happened to be present in all five vignettes. Yet, for a participant who thought the mystery word meant something else (e.g., *sheep*), there was no remembrance of horses past. Such a participant looks for a sheep but finds none present and posits a guess based on the current situation alone.

For the most part, we have used adults as proxies for all of the work described so far on the theory that information availability rather than conceptual sophistication is the right approach. But it would be reassuring, to say the least, if we could reproduce these results in young word learners. Woodard, Gleitman, and Trueswell (2015) did just this in a highly simplified laboratory study. Children ages 2 and 3 years "went on an animal safari" in which they viewed photographs of familiar and unfamiliar animals (see Figure 8). Critical trial sequences began with a presentation (Panel A) of an unfamiliar animal and a familiar one (e.g., a cow). Here only the familiar one was labeled ("Oh look a cow. Point to the cow.") On the next trial (Panel B) two additional unfamiliar animals are presented, and the spoken utterance "Oh look a mipen. Point to the mipen!" is heard. Children gaily pointed to one of the two animals without any basis for this choice. After two intervening filler trials in which known animals were labeled with their English names (Panel C), participants encountered the word "mipen" again. Here, children were assigned to one of two conditions. Children in the "No-Switch" condition viewed the animal they had selected previously alongside the unfamiliar animal that they had encountered earlier but never saw labelled (i.e., the animal from Panel A). Thus, for those children who had selected the animal on the right in Panel B, they were given the trial depicted in Panel D1. For those children who selected the left animal in Panel B, they were given the trial in Panel D2. Children in the "No Switch" condition showed very good retention of *mipen*, selecting the referent that they had selected before with 80% accuracy (Figure 9). Children in the "Switch" condition encountered a more malevolent referent world, one that mismatched their earlier conjecture: they viewed the animal they had not selected previously. So, those who had selected the animal on the right in Panel B were given Panel D2 whereas those who had selected the animal on the left in Panel B were given Panel D1. Our fundamental question is what happens under these conditions. A child who remembers the past situation of *mipen* (and its unselected alternative) should be well above chance. However, if all children can recall is their previous guess for *mipen*, they should now be at chance when faced with one animal never before

paired with *mipen* and another that had been but was unselected. Indeed, this was the case — they behave exactly as if this was the first time the word was encountered, a case of propose-but-failure-to-verify (Figure 9). Research from other laboratories provides supporting evidence from children of similar ages (Yurofsky & Frank, unpublished; Aravind, de Villiers, Pace, Valentine, Golinkoff, Hirsch-Pasek, Iglesias & Wilson, in press).

The learning patterns observed in Figure 7 and 9 were obtained from very different stimulus conditions (HSP vignettes and laboratory photographic images) and from subjects of different ages (adults and 2- and 3-year olds). Yet in both cases, subjects seem to store a single conjecture with a word form, discarding it only if it does not fit the succeeding referential context (i.e., it is not verified); in this latter case the subject seems to start again from scratch in finding the referent for the presently observed stimulus. From these data, we suggest learners do not aggregate situation-word-form pairs over many instances, seeking the best fit gradually and overall. Rather, learners do one trial learning with verification (one-and-a-bit trial learning). What this means is that assuming that the moment of correct identification varies across subjects, aggregation of accuracy across these subjects will look like gradual learning (a point made by several prior investigators of learning, including Gallistel, Balsam, & Fairhurst, 2004; Rock, 1957; Rodiger & Arnold, 2012). Indeed, we observe a gradual learning pattern of this sort - on average subjects as a group gradually get better, but each individual may learn at a different moment in the sequence (see Trueswell et al., 2013). Thus no individual “gradually learns” a word’s meaning, even though the group as a whole gradually improves.

3.2 Not so fast, the problem of homophones

Despite these successes, propose-but-verify contains the seeds of its own potential destruction: in principle, this procedure cannot acquire homophones. If the procedure had already proposed the interpretation *the-proboscis-of-an-elephant* it would expunge this conjecture upon next seeing a heavy piece of luggage, toggling back-and-forth between two meanings of “trunk.” The great virtue of propose-but-verify is that it immediately and irrevocably discards a large number of unlikely hypotheses (undetached rabbit parts, or luggage in the shape of an elephant’s nose, or a sheephorse for that matter). Its other great virtue is rapidly accepting the likely hypothesis upon passage of the simple checking procedure (i.e., verification). However, for homophones with two meanings of similar frequency, the checking procedure may not be effective, for the next occurrence of this word may be its evil homophonous twin, leading to rejection of the (correct) alternative meaning. In response to this, Stevens, Yang, Trueswell and Gleitman (2015) offered a friendly amendment to propose-but-verify that made it more robust to homophony and other conditions of noise - namely a disconfirmed hypothesis can be retained and later evaluated. Stevens et al. (2015) offered a computational implementation of this variant of propose-but-verify, which they called Pursuit. Pursuit, like propose-but-verify and unlike cross-situational models, stores only one referent per learning instance, but unlike propose-but-verify, a rejected hypothesis can return for consideration later in a learning sequence, leading to two entirely separate entries for the homophone. (This, by the way is consistent with the well known observation that young children who know both meanings of a homophone nevertheless fail to notice the ambiguity nor understand puns that play with these words,

Hirsh-Pasek, Gleitman & Gleitman, 1978). In a series of model comparisons, Stevens et al. found that Pursuit captured key findings in the literature better than either a state-of-the-art full-cross-situational model or the previous, more fragile, propose-but-verify learning procedure (see also Yang, this volume). Pursuit also performed as well or better than the full cross-situational model in overall measures of success when learning word meanings from corpora of child-directed speech. This latter demonstration, of a model using more limited data but performing as well or better than a model with full information, was explained as a case of limiting the detrimental effect of referentially ambiguous word contexts, junk.

3.3 Further Filters on Learning: A Return to Timing

Note that if learners *also* had access to information that permits the filtering (or down-weighting) of incorrect referential guesses, rare highly informative learning instances would on average have an even greater positive impact on the learner. Indeed, Experiment 2 of Trueswell et al. (2015) found that HSP observers are more confident about correct guesses as compared to incorrect ones. *This occurred despite the absence of explicit feedback, indicating a role for implicit feedback from the observed visuo-social context*. They identified one plausible candidate for such feedback: the timing of visuo-social cues to reference relative to word onset. As discussed above, observers were less confident about their correct guesses when the beep was surreptitiously offset from the actual occurrence of the word, even by just one second. Under this account, when the observer has a referential hypothesis in mind, expectations exist about how interactions with the referent object will unfold in time relative to the word's utterance. When these expectations are not met, confidence in that guess drops and a different guess may be posited. They proposed that word learners (children and adults) have implicit sensitivity to this timing information, and use it to determine if a referential hypothesis for an utterance is a good one. If the learner's referential hypothesis does not comport well with the behaviors of the speaker or of the target referent (e.g., if the hypothesized referent attracted attention too soon or too late) then this referential hypothesis is decremented. This proposal would suggest that although learners hear new words again and again (in this sense the stimulus situation is "cross-situational"), they likely attempt word learning only or primarily during rare single-situation events when cues to reference and their timing are satisfied. The learner who monitors and selects for precise temporal coupling between event and utterance is likely to experience occasional "epiphany moments" that push learning forward and dovetail with the observation that word learning is rapid, quite errorless (at least to the level of referent identification), and often occurs on a single or very few exposures.

What this means for early word learning is that, especially for situations of high ambiguity (junk), learners may be less likely to posit any referential hypothesis at all. Propose-but-verify and its later variants assumed that learners guessed a single hypothesis in the face of many alternatives (rather than storing them all). In this way, propose-but-verify could reduce the deleterious impact of ambiguity (it entered only a little bit of junk into learning rather than a lot). However, if instead, referential guesses are entered into learning only when there is reason to have confidence in that guess, some additional junk may be ignored entirely. Future research will need to explore the role of confidence within and across potential learning instances.

4. Conclusions

By general consensus, vocabulary acquisition is a matter of pairing sounds with their meanings - to a first approximation, hearing “dog” upon seeing a dog. However as we remarked in beginning, this kind of theory is a lot easier to assert than to defend, owing to the richness of the human conceptual and perceptual repertoire, and the buzzing-blooming complexity of the world that engages the system. Our general approach has been to find ways in which the scope of this problem is naturally reduced by the learner, in ways that account for, and hopefully explain, the course and character of child word learning. To these ends, we proposed a theory of vocabulary learning in 2005, that was based primarily on information likely to be available to the novice learner, which would change in both quality and kind during the first few years of life. Specifically we offered the following picture, restated directly from Gleitman et al. (2005):

1. Several sources of evidence contribute to solving the mapping problem for the lexicon.
2. These evidential sources vary in their informativeness over the lexicon as a whole.
3. Only one such evidential source is in place when word learning begins; namely, observation of the word’s situational contingencies.
4. Other systematic sources of evidence have to be built up by the learner through accumulating linguistic experience.
5. As the learner advances in knowledge of the language, these multiple sources of evidence converge on the meanings of new words. These procedures mitigate and sometimes reverse the distinction between “easy” and “hard” words.

We accepted, to avoid the problem of circularity, that the anchor point of this procedure was to be a small set of words, acquired from observation without collateral linguistic evidence, much as proposed by Grimshaw (1994) and Pinker (1984) and documented in vocabulary growth findings from Hart & Risley 1995 and others (see Figure 1). The present essay has returned to ask in detail whether this default assumption is tenable and how it might work in detail.

The first half of this essay showed that the complications and richness of the raw input are real and raise the question of an effective procedure for dealing with it. One classical and modern theme of the solution for solving problems in a noisy environment has been the use of “big data”. Indeed the intelligent machine that engages in a vast aggregation process can produce incredibly impressive results on language data (e.g., Dyer et al., 2015; Hinton et al., 2012; Wu et al., 2016). And that is very much in tune with the kinds of associative learning procedures that have been considered since the British Empiricists and in the tradition of Structural Linguistics (Harris, 1964). Yet there are many reasons to wonder about such theories as analogs for how the human mind works. One of them is that, at least on first inspection, children appear to learn a word “in a flash” based on one or a very few experiences rather than thousands. We have tried to convince you that child word learners are not considering many hypotheses from a single learning instance, but rather find ways to

silence the noise in the data with a series of perceptual and procedural filters. In terms of perceptual filters, we have emphasized timing, the exact temporal confluence of word with world. In terms of procedural filters, we have suggested the learner extracts one hypothesis at a time, seeking verification.

Supposing our informal model of primitive word learning to be correct in its general properties, the explanatory victory seems at first glance to be hollow because, as we acknowledged, word-to-world pairing is a crude, domain-general procedure which can account for perhaps 1 or 2 percent of the vocabulary present in adulthood (see Carey, 1978; Bloom, 2000). At best, it describes only the first few hundred words understood and lisped out by children in the first year and a half of life. Not only is this vocabulary stage brief, it is quite restricted in its types, comprised mainly of whole-object nouns (see Gentner 1982; Gentner & Boroditsky, 2001) and a smattering of other words which are very likely to occur just when the relevant event is happening (“all gone” when the plate is clean, “up” as the child is being lifted, “uh-oh” when an error or accident is happening).

Nonetheless the primitive vocabulary thus achieved is central to understanding language acquisition for four main reasons. First, the social intercourse between adults and infants is materially advanced when even a little mutually-known vocabulary is present (see, e.g., Golinkoff, 1986). Second, this primitive vocabulary draws solely on the sophisticated knowledge of perceptual and belief-desire psychology that infants bring into the learning situation (e.g., Gergely, Bekkering & Kiraly, 2002; Leslie, 1987; Leslie & Keeble, 1987; Spelke, 1990; Baillargeon, Scott & He, 2010); that is, infants understand “the world” though not the exposure language.

Third, the primitive vocabulary allows learners to uncover the specifics of their language’s syntax insofar as this varies (cf. Gleitman et al. 2005 among many other sources). In particular, the phrase structure representation enabled by the easy words makes possible the reconstruction of the argument structure of sentences (“whole thoughts”). This can’t be done absent specific language knowledge because, both within and across languages, arguments can be dropped (e.g., Mandarin Chinese versus French), oblique cases vary in their marking in surface sentences (e.g., Finnish versus English), and sentential Subjects can appear either serially early or late in the basic language structures (e.g., English vs Fijian). To organize this variable though constrained database, the easy words inserted intonationally into the stream of words, can do much work. This change in the representational power of the input analysis - from word-to-world mapping to structure-to-world mapping - can explain how we distinguish “hugging” from “giving” (argument number, see Fisher, 1996; Gordon, 2003; Lidz et al, 2003; Naigles, 1990), when it is better to ‘receive’ than to ‘give’ (argument position; Fisher, Hall, Rakowitz & Gleitman, 1994.; Nappa et al., 2009), and, perhaps most interesting of all, how the blind learn “to see” (Landau & Gleitman, 1985).

The fourth virtue of our position is its potential for explaining two broad characteristics of the early learning procedure: why it is so slow and fast at the same time. It is slow in regard to vocabulary accrual in which the first words (mainly, the first 200 or so words produced and understood) are acquired at the rate of less than a word per week (Hart & Risley, 1995). We explain this slowness, as well as the types of words then acquired (largely, concrete

terms), because the information comes solely from observing how words line up with their conversational contingencies. A sudden increase in the word learning rate occurs at the moment when children bring new sources of evidence on line, taking into account not only the word as an isolated element, but as rendered within a syntactic structure (Lenneberg, 1967; Landau & Gleitman, 1985, inter alia). In contrast, word learning is fast from the beginning in another sense: rather than requiring hundreds or thousands of exposures to acquire the easy words, infants seek out and acquire them at auspicious moments when the likelihood of some single hypothesis has been maximized. We have tried to explain the efficiency of this process in terms of a variety of filters, conceptual, perceptual, and procedural, that sort the gems from the junk.

Acknowledgements

We would like to thank the many collaborators who contributed to research presented here and the many discussions extending over years at the Cheese and Trackers Research seminar. We thank Victor Gomes for help with proofreading of the paper. Research supported by National Institute of Child Health and Human Development Grant R01HD37507 (P.I.s: L.G. and J.T.).

References

Aravind A de Villiers J, Pace A, Valentine H, Golinkoff R, Hirsch-Pasek K, Iglesias A & Wilson M (submitted). Fast mapping word meanings across trials: young children forget all but their first guess. Paper submitted for publication.

Baillargeon R, Scott RM, & He Z (2010). False-belief understanding in infants. Trends in Cognitive Sciences, 14(3), 110–118. [PubMed: 20106714]

Baldwin DA (1991). Infants' contribution to the achievement of joint reference. Child Development, 62(5), 874–890.

Baldwin DA (1993). Infants' ability to consult the speaker for clues to word reference. Journal of Child Language, 20, 395–418. [PubMed: 8376476]

Baldwin DA (1995). Understanding the link between joint attention and language In Moore C & Dunham PJ (Eds.), Joint Attention: Its origins and role in development (pp. 131–158). Hillsdale, NJ, England: Lawrence Eribaum Associates, Inc.

Bates E, Dale PS, & Thal D (1995). Individual differences and their implications for theories of language development In Fletcher P, & MacWhinney B, (Eds.) Handbook of Child Language, (pp. 96–151). Oxford: Basil Blackwell.

Bloom P (2000). How Children Learn the Meanings of Words. Cambridge, MA: MIT Press.

Bruner J (1974/1975). From communication to language - A psychological perspective. Cognition, 3(3), 255–287

Carey S (1978). The child as word learner In Halle M, Miller G, & Bresnan J (Eds.), Linguistic Theory and Psychological Reality (pp. 264–293). Cambridge, MA: MIT Press.

Caselli MC, Bates E, Casadio P, Fenson J, Fenson L, Sanderl L, & Weir J (1995). A cross-linguistic study of early lexical development. Cognitive Development, 10(2), 159–199.

Cartmill EA, Armstrong BF, Gleitman LR, Goldin-Meadow S, Medina TN, & Trueswell JC (2013). Quality of early parent input predicts child vocabulary 3 years later. Proceedings of the National Academy of Sciences, 110(28), 11278–11283.

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

Clerkin EM, Hart E, Rehg JM, Yu C, & Smith LB (2017). Real-world visual statistics and infants' first-learned object names. Phil. Trans. R. Soc. B, 372(1711), 20160055. [PubMed: 27872373]

Doherty MJ, Anderson JR & Howeieson L (2009). The rapid development of explicit gaze judgment ability at 3 years. Journal of Experimental Child Psychology, 104, 296–312. [PubMed: 19640550]

Dyer C, Ballesteros M, Ling W, Matthews A, & Smith N (2015).Transition-based dependency parsing with stack long short-term memory. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics, 334–343.

Elman JL (1993). Learning and development in neural networks: The importance of starting small. Cognition, 48(1), 71–99. [PubMed: 8403835]

Fisher C (1996). Structural limits on verb mapping: The role of analogy in children's interpretations of sentences. Cognitive Psychology, 31(1), 41–81. [PubMed: 8812021]

Fisher C, & Gleitman LR (2002). Language acquisition In Pashler HF (Series Ed.) and Gallistel CR (Volume Ed.), Stevens' Handbook of Experimental Psychology, Vol 1: Learning and Motivation (pp. 445–496). New York: Wiley.

Fisher C, Gleitman H, & Gleitman LR (1991). On the semantic content of subcategorization frames. Cognitive Psychology, 23(3), 331–392. [PubMed: 1884596]

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

Fodor JA (1975). The Language of Thought. Harvard University Press.

Frege G (1892/1948). Sense and reference. The Philosophical Review, 57(3), 209–230.

Gallistel CR, Balsam PD, & Fairhurst S (2004). The learning curve: Implications of a quantitative analysis. Proceedings of the National Academy of Sciences USA, 101(36), 13124–13131.

Gentner D (1982). Why nouns are learned before verbs: linguistic relativity versus natural partitioning In Kuczaj SA II (ed .), Language development, Vol. 2: Language, thought, and culture. Hillsdale, NJ: Lawrence Erlbaum, 301–3 3

Gentner D, & Boroditsky L (2001). Individuation, relativity and early word learning In Bowerman M & Levinson S (Eds.), Language acquisition and conceptual development (pp. 215–256). Cambridge, UK: Cambridge University Press.

Gergely G, Bekkering H, & Kiraly I (2002). Rational imitation in preverbal infants. Nature, 415, 755.

Gillette J, Gleitman H, Gleitman L, & Lederer. (1999). Human simulations of vocabulary learning. Cognition, 73, 135–176. [PubMed: 10580161]

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

Gleitman LR, Cassidy K, Nappa R, Papafragou A, & Trueswell JC (2005). Hard Words. Language Learning and Development, 1(1), 23–64.

Gleitman L, & Fisher C (2005). Universal aspects of word learning In McGilvray JA (Ed.) The Cambridge Companion to Chomsky (pp. 123–142). Ernst Klett Sprachen.

Gleitman LR, & Landau B (Eds.). (1994). The Acquisition of the Lexicon. MIT Press.

Goldin-Meadow S, Levine SC, Hedges LV, Huttenlocher J, Raudenbush S, & Small S (2014). New evidence about language and cognitive development based on a longitudinal study: Hypotheses for intervention. American Psychologist, 69(6), 588–599. [PubMed: 24911049]

Golinkoff RM (1986). "I beg your pardon?": The preverbal negotiation of failed messages. Journal of Child Language, 13, 455–476. [PubMed: 3793809]

Gordon P (2003). The origin of argument structure in infant event representations. In Proceedings of the 26th Boston University Conference on Language Development (pp. 189–198).

Grimshaw J (1994). Lexical reconciliation In Gleitman LR & Landau B (Eds.), The acquisition of the lexicon. Cambridge, MA: MIT Press.

Harris ZS (1964). Co-occurrence and Transformation in Linguistic Structure. Englewood Cliffs, NJ: Prentice-Hall.

Hart B, & Risley TR (1995). Meaningful differences in the everyday experience of young American children. Baltimore, MD: Paul H. Brookes Publishing.

Harris M, Jones D, Brookes S, & Grant J (1986). Relations between the non-verbal context of maternal speech and rate of language development. British Journal of Developmental Psychology, 4(3), 261–268.

Hinton G Deng L, Yu D, Dahl G et al., (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. Signal Processing Magazine, 29, 82–97.

Hirsh-Pasek K, Gleitman LR, & Gleitman H (1978). What did the brain say to the mind? A study of the detection and report of ambiguity by young children In The child's conception of language (pp. 97–132). Springer, Berlin, Heidelberg.

Hume D (1748/1902). An Enquiry Concerning Human Understanding In Selby-Bigge LA (Ed.), An Enquiry Concerning Human Understanding and Concerning the Principles of Morals. Oxford: Clarendon Press (Original work published in 1748)

Jaswal VK (2010). Believing what you're told: Young children's trust in unexpected testimony about the physical world. Cognitive Psychology, 61, 248–272. [PubMed: 20650449]

Johnson EK, & Jusczyk PW (2001). Word segmentation by 8-month-olds: When speech cues count more than statistics. Journal of Memory and Language, 44(4), 548–567.

Jusczyk PW, & Hohne EA (1997). Infants' memory for spoken words. Science, 277, 1984–1986. [PubMed: 9302291]

Katz JJ, & Fodor JA (1963). The structure of a semantic theory. Language, 39(2), 170–210.

Landau B, & Gleitman LR (1985). Language and Experience: Evidence from the Blind Child. Cambridge, MA: Harvard University Press.

Lenneberg EH (1967). Biological foundations of language. New York: Wiley

Leslie AM (1987). Pretense and representation: The origins of "theory of mind". Psychological Review, 94(4), 412–426.

Leslie AM, & Keeble S (1987). Do six-month old infants perceive causality? Cognition, 25, 265–288. [PubMed: 3581732]

Lidz J, Gleitman H, & Gleitman L (2003). Understanding how input matters: Verb learning and the footprint of universal grammar. Cognition, 87(3), 151–178. [PubMed: 12684198]

Locke J (1690). An Essay Concerning Human Understanding.

Marcus GF, Vijayan S, Rao SB, & Vishton PM (1999). Rule learning by seven-month-old infants. Science, 283(5398), 77–80. [PubMed: 9872745]

Medina TN, Snedeker J, Trueswell JC, & Gleitman LR (2011). How words can and cannot be learned by observation. Proceedings of the National Academy of Sciences, 108, 9014–9019.

Michotte A (1963). The perception of causality. New York: Basic Books (Original work published 1946)

Naigles L (1990). Children use syntax to learn verb meanings. Journal of Child Language, 17(2), 357–374. [PubMed: 2380274]

Naigles LR (1996). The use of multiple frames in verb learning via syntactic bootstrapping. Cognition, 58(2), 221–251. [PubMed: 8820388]

Naigles L, Gleitman LR, & Gleitman H (1986). Children acquire word meaning components from syntactic evidence Language and Cognition: A Developmental Perspective. Norwood, NJ: Ablex.

Nappa R, Wessell A, McEldoon KL, Gleitman LR, & Trueswell JC (2009). Use of speaker's gaze and syntax in verb learning. Language Learning and Development, 5(4), 203–234. [PubMed: 24465183]

Newport EL (1990). Maturational constraints on language learning. Cognitive Science, 14, 11–28.

Papafragou A, Friedberg C, & Cohen ML (2017). The role of speaker knowledge in children's pragmatic Inferences. Child Development.

Papafragou A, Massey C, & Gleitman L (2006). When English proposes what Greek presupposes: The cross-linguistic encoding of motion events. Cognition, 98(3), B75–B87. [PubMed: 16043167]

Piccin TB, & Waxman SR (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.

Pinker S (1984). Language Learnability and Language Development. Cambridge, MA: Harvard University Press.

Pinker S (1994). How could a child use verb syntax to learn verb semantics? Lingua: International Review of General Linguistics, 92(1–4), 377–410.

Quine WVO (1960). Word and object (Studies in Communication). New York and London: Technology Press of MIT.

Rock I (1957). The role of repetition in associative learning. American Journal of Psychology, 70, 186–193. [PubMed: 13424758]

Roediger HL III, & Arnold KM (2012). The one-trial learning controversy and its aftermath: Remembering Rock (1957). American Journal of Psychology, 125(2), 127–143. [PubMed: 22774677]

Roy BC, Frank MC, & Roy D (2009). Exploring word learning in a high-density longitudinal corpus.

Rosch E (1975). Cognitive representations of semantic categories. Journal of Experimental Psychology: General, 104(3), 192.

Rosch E, & Mervis CB (1975). Family resemblances: Studies in the internal structure of categories. Cognitive Psychology, 7(4), 573–605.

Saffran JR, Aslin RN, & Newport EL (1996). Statistical learning by 8-month-old infants. Science, 274(5294), 1926–1928. [PubMed: 8943209]

Siskind JM (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. Cognition, 61(1–2), 39–91. [PubMed: 8990968]

Slobin DI (2008). The child learns to think for speaking: Puzzles of crosslinguistic diversity in form-meaning mappings In Ogura T, Kobayashi H, et al. (Eds.), Studies in Language Sciences 7, 3–22. Tokyo: Kurosio Publishers.

Smiley P, & Huttenlocher JE (1995). Conceptual development and the child's early words for events, objects, and persons In Tomasello M & Merriman W, Beyond Names for Things (pp. 21–62). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Smith LB, Colunga E, & Yoshida H (2010). Knowledge as process: contextually cued attention and early word learning. Cognitive Science, 34(7), 1287–1314. [PubMed: 21116438]

Smith LB, Yu C, & Pereira AF (2011). Not your mother's view: The dynamics of toddler visual experience. Developmental Science, 14(1), 9–17. [PubMed: 21159083]

Snedeker J, & Gleitman LR (2003). Why it is hard to label our concepts In Waxman S & Hall G (Eds.), Weaving a Lexicon. NY: Cambridge University Press.

Southgate V, Chevallier C, & Csibra G (2010). Seventeen-month-olds appeal to false beliefs to interpret others' referential communication. Developmental Science, 13(6), 907–912. [PubMed: 20977561]

Spelke ES (1990). Principles of object perception. Cognitive Science, 14(1), 29–56.

Stevens JS, Gleitman LR, Trueswell JC, & Yang C (2017). The pursuit of word meanings. Cognitive Science, 41(4), 638–676. [PubMed: 27666335]

Thiessen ED, & Saffran JR (2003). When cues collide: use of stress and statistical cues to word boundaries by 7-to 9-month-old infants. Developmental Psychology, 39(4), 706. [PubMed: 12859124]

Tomasello M (1995). Pragmatic contexts for early verb learning In Tomasello M & Merriman W (Eds.), Beyond Names for Things: Young Children's Acquisition of Verbs. Lawrence Erlbaum.

Tomasello M, & Farrar MJ (1986). Joint attention and language. Child Development, 57(6), 1451–1463.

Tomasello M, Mannle S, & Kruger AC (1986). Linguistic environment of 1- to 2-year-old twins. Developmental Psychology, 22(2), 169–176.

Tomasello M, & Todd J (1983). Joint attention and lexical acquisition style. First Language, 12, 197–211.

Trueswell JC, Dawson T, Pozzan L & Gleitman LR (in prep.). Referential Timing in Word Learning. Manuscript in preparation.

Trueswell JC, Lin Y, Armstrong B, Cartmill EA, Goldin-Meadow S, & Gleitman LR (2016). Perceiving referential intent: Dynamics of reference in natural parent-child interactions. Cognition, 148, 117–135. [PubMed: 26775159]

Turing AM (1950). Computing machinery and intelligence. Mind, 59(236), 433–460.

Wexler K, & Culicover P. (1980). Formal Principles of Language Acquisition. Cambridge, MA: MIT Press.

Woodard K, Gleitman LR, & Trueswell JC (2016). Two-and three-year-olds track a single meaning during word learning: Evidence for Propose-but-verify. Language Learning and Development, 12(3), 252–261. [PubMed: 27672354]

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

Wu Y, Schuster M, Chen Z, Le QV, Norouzi M, Macherey W et al. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. ArXiv E-prints, 2arXiv:1609.08144.

Yu C, & Smith LB (2007). Rapid word learning under uncertainty via cross-situational statistics. Psychological Science, 18(5), 414–420. [PubMed: 17576281]

Yuan S, Fisher C, & Snedeker J (2012). Counting the nouns: Simple structural cues to verb meaning. Child Development, 83(4), 1382–1399. [PubMed: 22616898]

Yurovsky D, Smith LB, & Yu C (2013). Statistical word learning at scale: The baby's view is better. Developmental Science, 16(6), 959–966. [PubMed: 24118720]

Yurofsky D & Frank M (2013). Active hypothesis testing and co-occurrence tracking work together in cross-situational word learning. Paper presented at the 38th Annual Boston University Conference on Language Development.

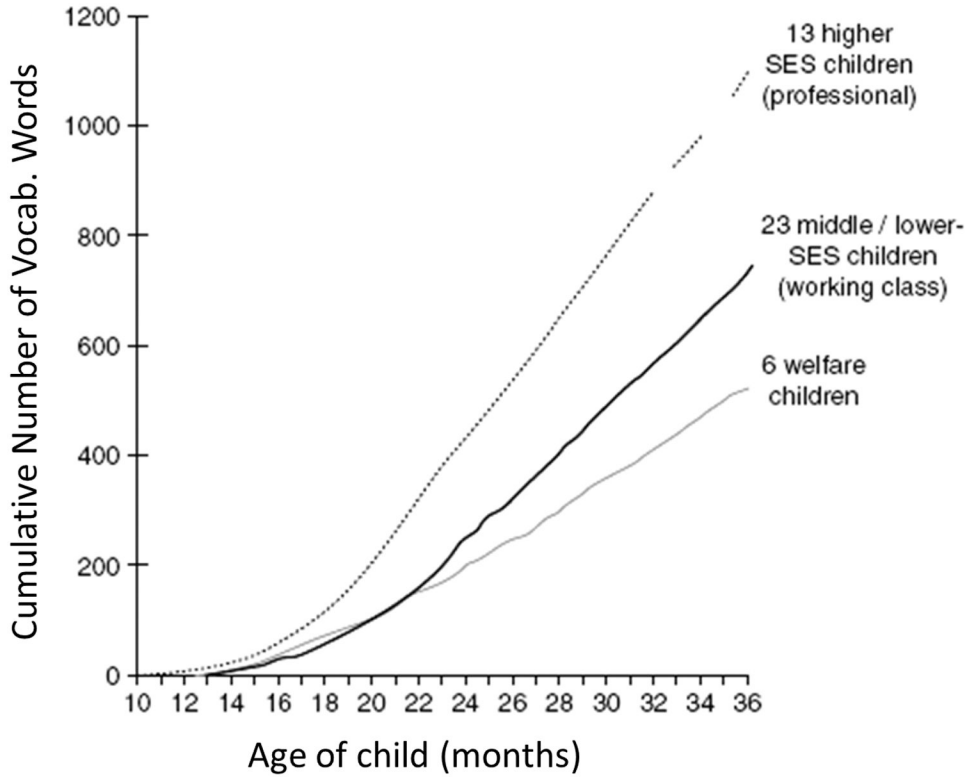

Figure 1.

Number of word types produced as a function of age of child. Figure adapted from *Meaningful differences in the everyday experience of young American children* (p. 234), B. Hart and T. Risley, 1995, Paul H Brookes Publishing.

Figure 2.
Example of a referential context. Figure reproduced from Medina et al. (2011). Copyright 2011, *Proceedings of the National Academy of Sciences*.

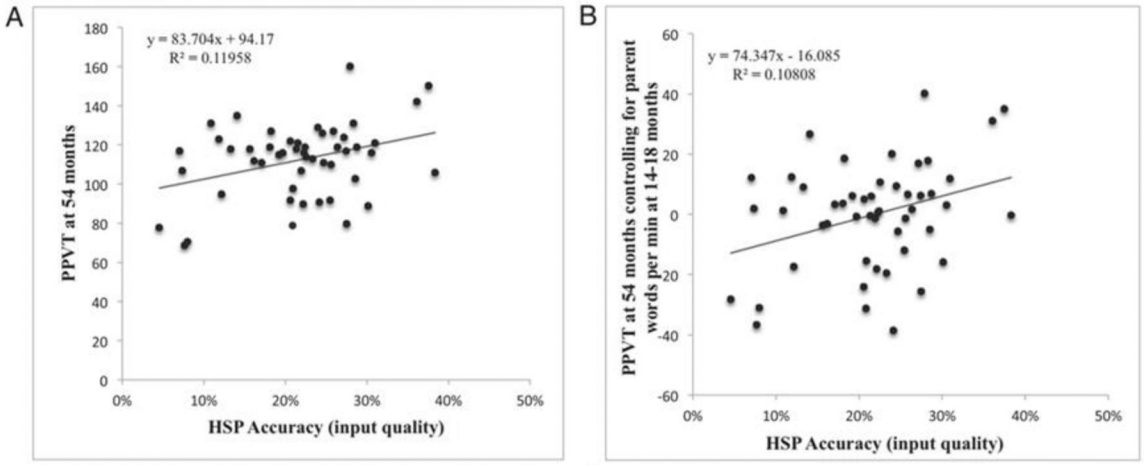

Figure 3. The Quality of Referential Environment in the Home Predicts Child Vocabulary Size Three Years Later.

Each data point represents a parent-child dyad. The x-axis is the average accuracy with which naive adult observers could guess what the parent in this dyad was saying to their 14–18 month old child, based on a set of muted video vignettes of everyday parent-child interactions in the home at age 14–18 months (Human-Simulation-Paradigm, HSP, Accuracy). The y-axis is that same child's vocabulary size three years later at school entry (Peabody Picture Vocabulary Test, PPVT, at age 54 months). Panel A is the direct relationship, and Panel B is the relationship after controlling for the quantity of early input (the average number of words per minute uttered by the parent to their child at 14–18 months). Figure reproduced from Cartmill et al. (2013). Copyright 2013, *Proceedings of the National Academy of Sciences*.

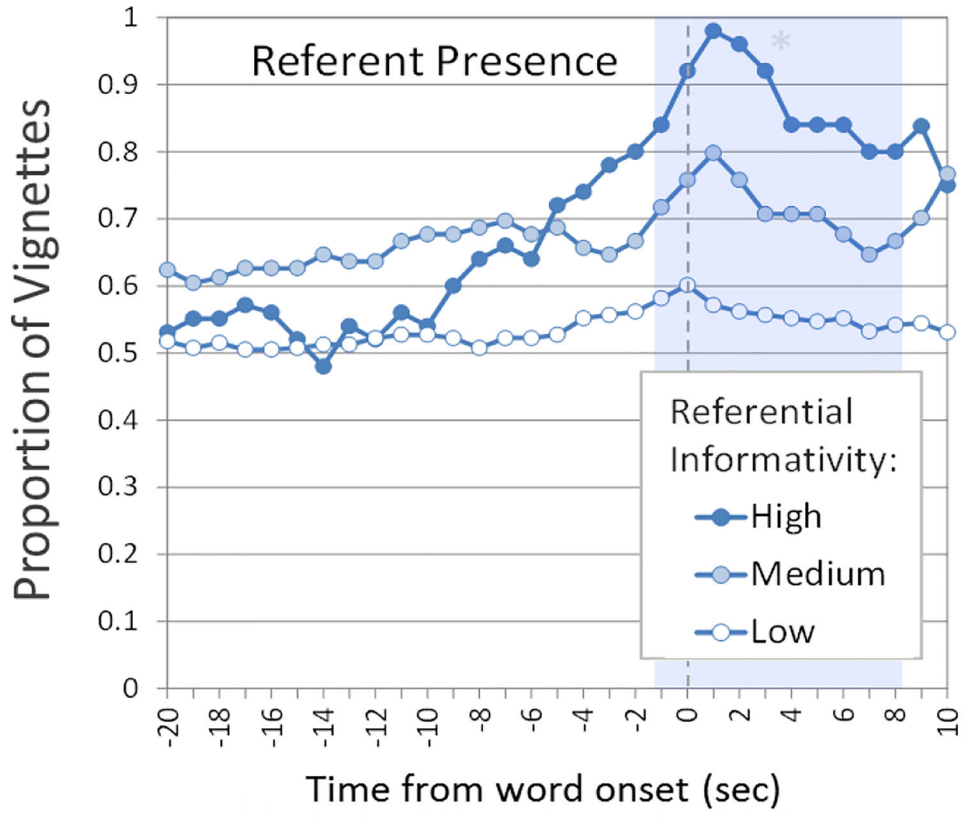

Figure 4. Object appearance, not continuous presence characterizes high informative contexts.
Proportion of Human-Simulation-Paradigm (HSP) Vignettes in which the object being referred to the parent was coded as present, on a second-by-second basis. High Informative Vignettes were ones where HSP observers guessed the parent's utterance over 50% of the time. These vignettes were rare and characterized by an increased probability of the referent being present just before word onset, peaking at essentially 100% presence 1 second after. Low Informative Vignettes were those where HSP observers guessed correctly less than 10% of the time. These vignettes were common, had the referent present on about 50% of vignettes, with no changes over time. Medium vignettes were the rest of all vignettes and fell in between. Shaded areas reflect the time periods for which the social-attentive behavior was reliably predicted by HSP accuracy scores (* $p<0.05$). Adapted from Trueswell et al. (2016). *Elsevier Press*.

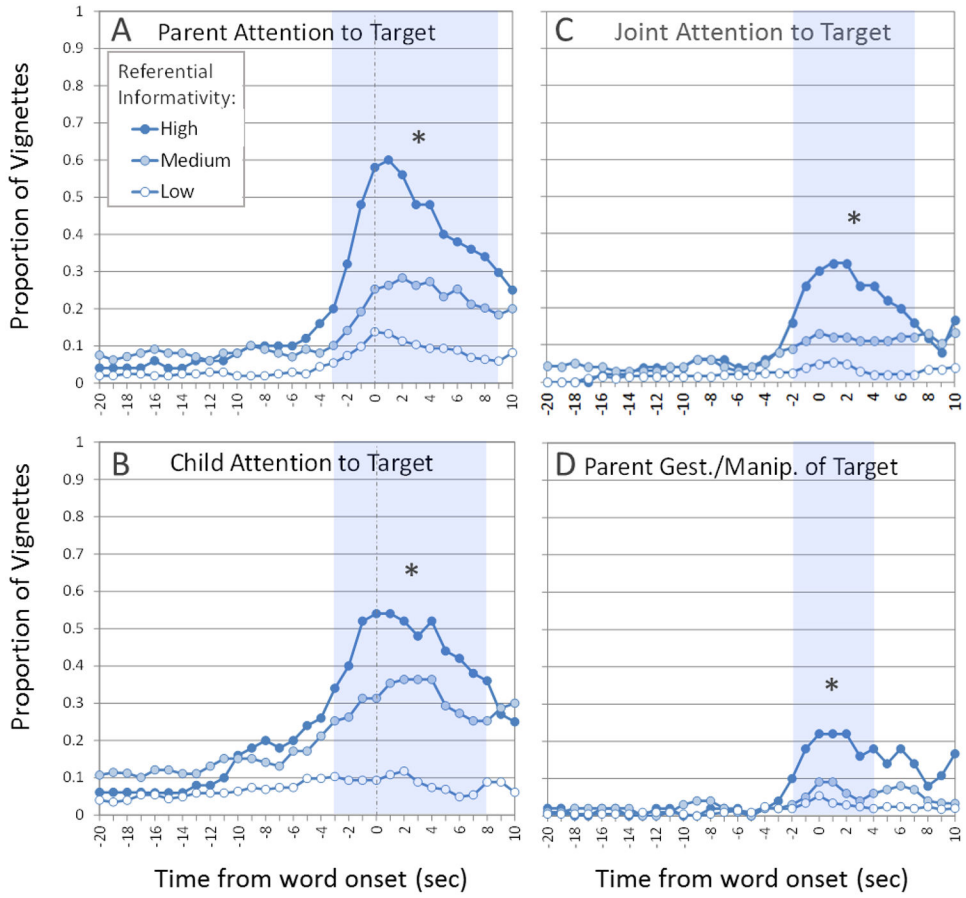

Figure 5.

Proportion of Human-Simulation-Paradigm (HSP) Vignettes that were coded as containing: (A) Parent attending to the referent; (B) Child attending to the referent; (C) The intersection of these two, i.e., joint attention; and (D) Parent gesturing at or manipulating the referent. High Informative Vignettes were ones where HSP observers guessed the parent's utterance over 50% of the time. These vignettes were rare and contained a sudden increase in the relevant social-attentive behaviors just prior to word onset. Low Informative Vignettes were those where HSP observers guessed correctly less than 10% of the time. These vignettes were common and lacked relevant social-attentive behaviors. Medium vignettes were the rest of all vignettes and fell in between. Shaded areas reflect the time periods for which the social-attentive behavior was reliably predicted by HSP accuracy scores (* $p<0.05$). Adapted from Trueswell et al. (2016). Copyright 2016, *Elsevier Press*.

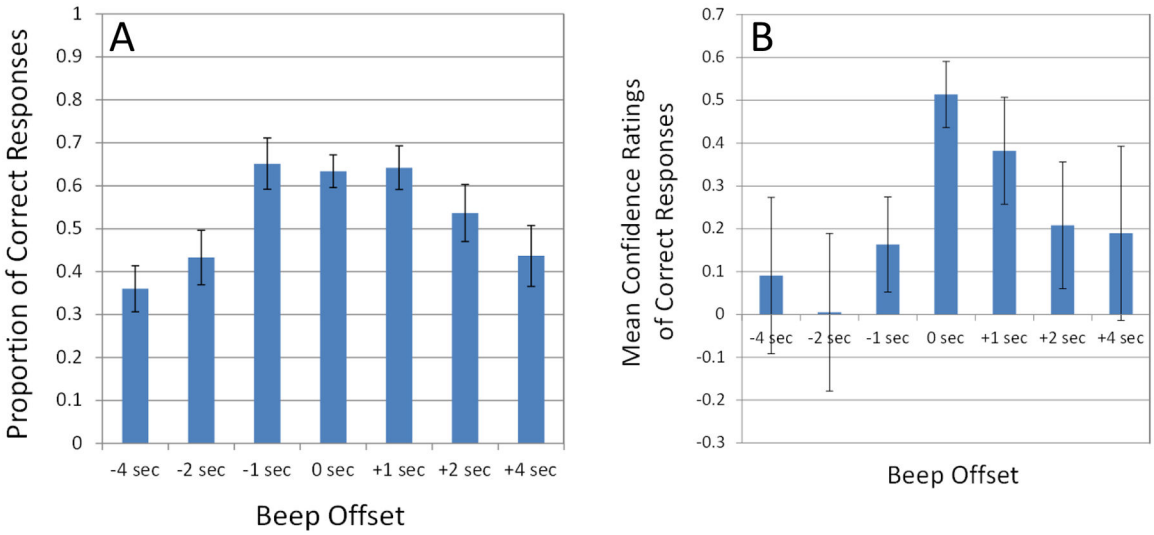

Figure 6.

A. Average proportion of correct responses (HSP accuracy) as a function of beep offset from actual word occurrence. B. Average normalized confidence ratings (z-scores within each subject) for correct HSP responses only as a function of beep offset from actual word occurrence. Averages based on subject means. Error bars indicate 95% confidence intervals. Reproduced from Trueswell et al. (2016). *Copyright Elsevier Press*.

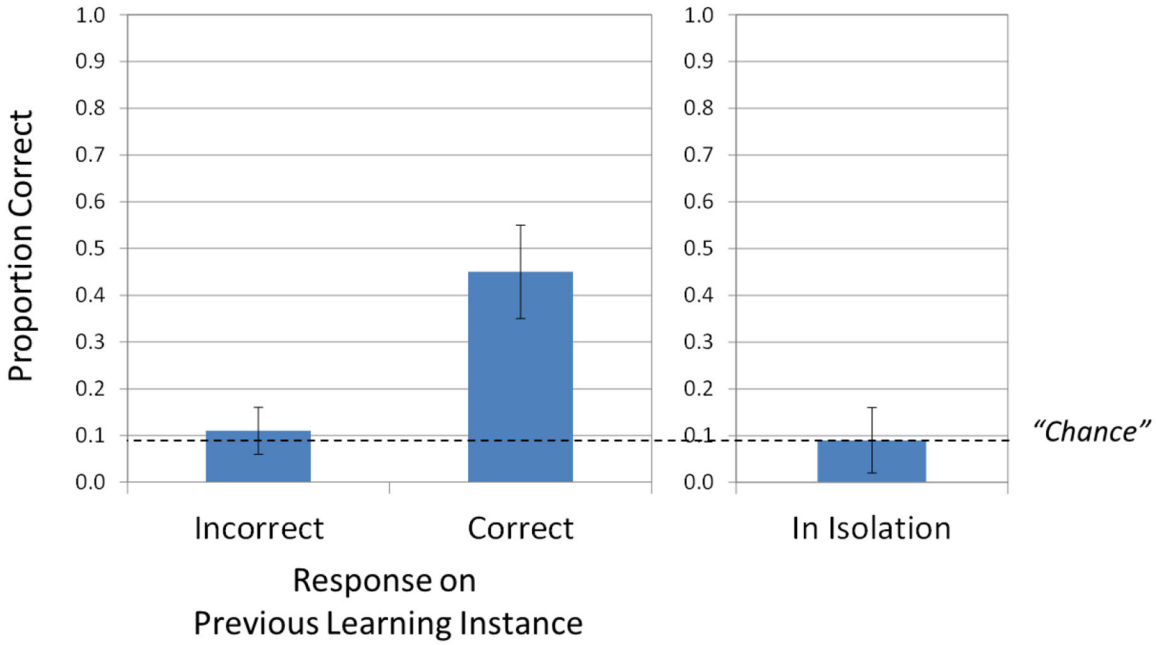

Figure 7. People recall only their previous hypothesis for a word's meaning, not the previous situational context.

When people made an incorrect guess as to the referent to the word, performance on the next word's occurrence was no better than participants who had viewed the same vignette in isolation without the benefit of cross-situational comparison.

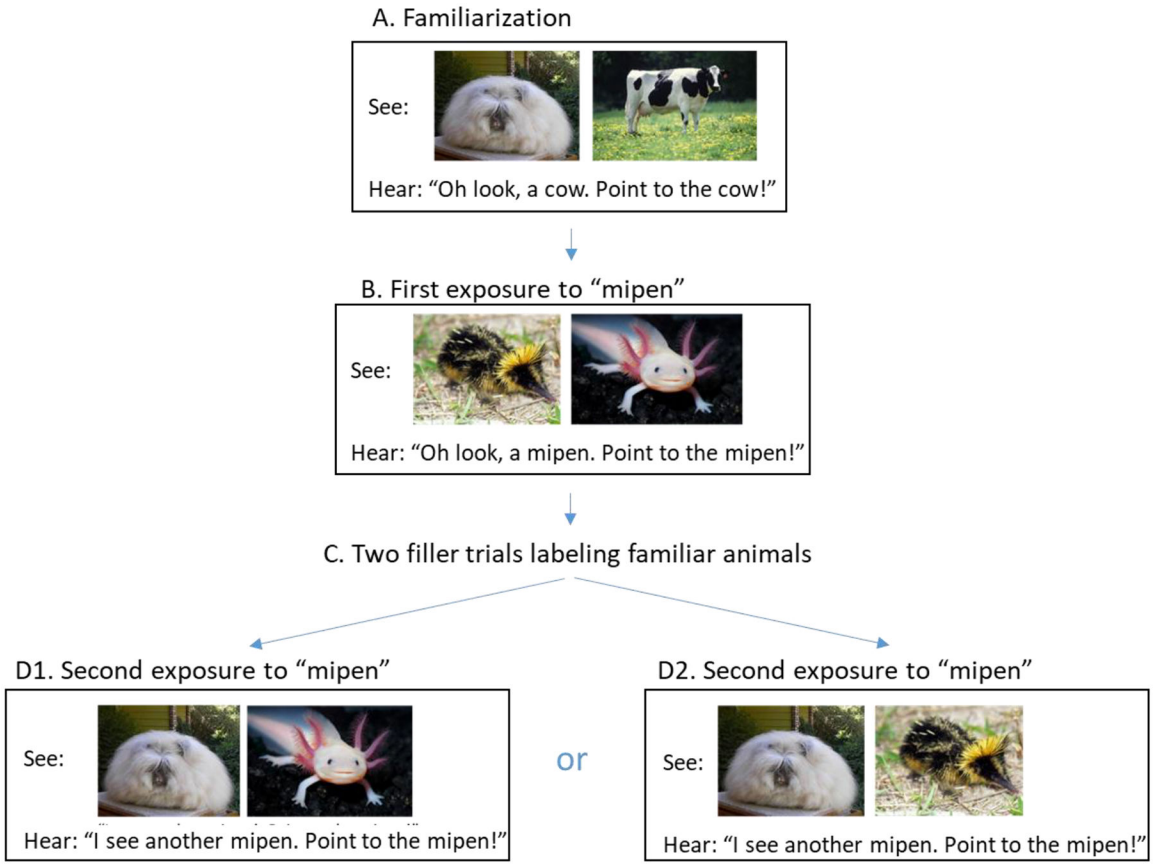

Figure 8. Trial sequence for test items.

Panel A. The child saw an unknown and known animal and was asked to point to the known animal. Panel B. The child then viewed two additional unknown animals and was asked to point to the “mipen”. Panel C. The child then encountered two filler trials involving known animals (no additional unknown animals were encountered). The child then encountered “mipen” again in either Panel D1 or D2. For a child assigned to the “No Switch” condition, they were shown the animal that they had selected previously. Thus, a child who selected the animal on the right in Panel C would then proceed to D1 whereas a child who selected the animal on the left in Panel C would proceed to D2. In contrast, a child in the “Switch” condition was shown the animal that they had not selected previously. Thus, a child who selected the animal on the right in Panel C would then proceed to D2 whereas a child who selected the animal on the left in Panel C would proceed to D1. Position of animals was randomized and counterbalanced across conditions.

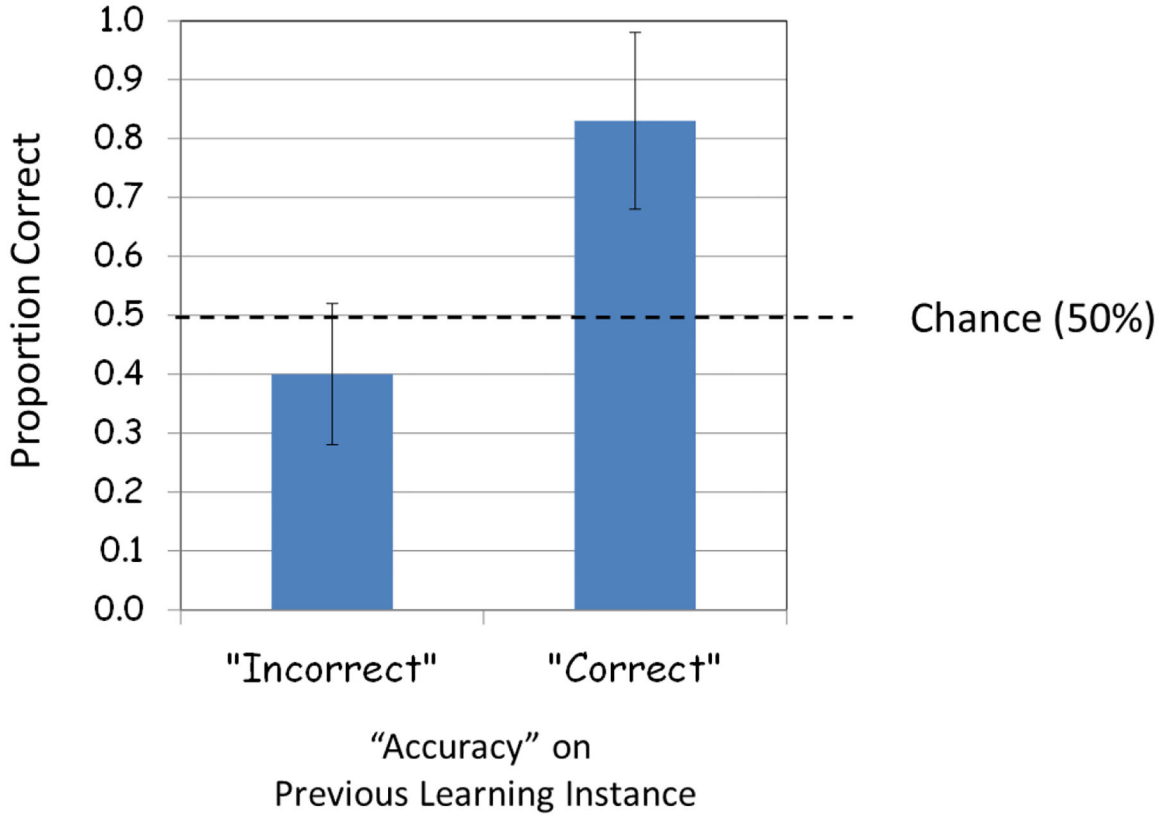

Figure 9. Two and three year olds propose-but-verify.

Young children, even under highly simplified laboratory contexts (Figure 8) recall only what the word referred to previously, not the context in which the word appeared. Children who were previously "correct" (the "No Switch" condition) selected the animal that had co-occurred with the nonsense word. Participants who were previously "incorrect" (the "Switch" condition) were at chance, suggesting that they did not recall that the animal had co-occurred with the nonsense word. These findings are very similar to the behavior of adults who were attempting to learn words from complex natural scenes (i.e., HSP vignettes, Figure 7) or from laboratory conditions that were referentially complex (Trueswell et al., 2013). Reproduced from Woodard et al. (2015). *Copyright Routledge Press.*