Cognition 148 (2016) 117–135

Contents lists available at ScienceDirect

Cognition

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

Perceiving referential intent: Dynamics of reference in natural parent–child interactions

John C. Trueswell [a,*], Yi Lin [a], Benjamin Armstrong III [a], Erica A. Cartmill [b], Susan Goldin-Meadow [c], Lila R. Gleitman [a,*]

[a] University of Pennsylvania, United States
[b] University of California, Los Angeles, United States
[c] University of Chicago, United States

ARTICLE INFO

Article history:
Received 3 August 2014
Revised 5 August 2015
Accepted 6 November 2015
Available online 8 January 2016

Keywords:
Psycholinguistics
Language development
Word learning
Reference

ABSTRACT

Two studies are presented which examined the temporal dynamics of the social-attentive behaviors that co-occur with referent identification during natural parent–child interactions in the home. Study 1 focused on 6.2 h of videos of 56 parents interacting during everyday activities with their 14–18 month-olds, during which parents uttered common nouns as parts of spontaneously occurring utterances. Trained coders recorded, on a second-by-second basis, parent and child attentional behaviors relevant to reference in the period (40 s) immediately surrounding parental naming. The referential transparency of each interaction was independently assessed by having naïve adult participants guess what word the parent had uttered in these video segments, but with the audio turned off, forcing them to use only non-linguistic evidence available in the ongoing stream of events. We found a great deal of ambiguity in the input along with a few potent moments of word-referent transparency; these transparent moments have a particular temporal signature with respect to parent and child attentive behavior: it was the object's appearance and/or the fact that it captured parent/child attention at the moment the word was uttered, not the presence of the object throughout the video, that predicted observers' accuracy. Study 2 experimentally investigated the precision of the timing relation, and whether it has an effect on observer accuracy, by disrupting the timing between when the word was uttered and the behaviors present in the videos as they were originally recorded. Disrupting timing by only ±1 to 2 s reduced participant confidence and significantly decreased their accuracy in word identification. The results enhance an expanding literature on how dyadic attentional factors can influence early vocabulary growth. By hypothesis, this kind of time-sensitive data-selection process operates as a filter on input, removing many extraneous and ill-supported word-meaning hypotheses from consideration during children's early vocabulary learning.

1. Introduction

Our intuitions tell us that infants likely learn the meanings of their very first words during moments when word and object happen to co-occur, e.g., when they hear the word "doggie" in the presence of a dog. And indeed, ample observational and experimental evidence supports this idea (e.g., Baldwin, 1991, 1993; Baldwin & Tomasello, 1998; Bloom, 2002; Brown, 1973; Hollich et al., 2000; Pruden, Hirsh-Pasek, Golinkoff, & Hennon, 2006; Smith, Colunga, & Yoshida, 2010). Yet this very same evidence tells us that mutual co-presence of word and thing is probabilistic and conditional, rather than necessary and sufficient, for an infant to identify a referent and learn a word's meaning. The referential context depicted in Fig. 1 is an example of one glaring problem that must be solved to make good on any word-to-referent scheme for lexical learning: there seem to be far too many hypotheses made available by the observed scene, and probably too many for a realistic full cross-situational comparison process to parse out across multiple observations (e.g., Medina, Snedeker,

* Corresponding authors at: Department of Psychology, University of Pennsylvania, 3720 Walnut Street, Solomon Lab Bldg., Philadelphia, PA 19104-6241, United States.
E-mail addresses: trueswel@psych.upenn.edu (J.C. Trueswell), gleitman@psych.upenn.edu (L.R. Gleitman).

http://dx.doi.org/10.1016/j.cognition.2015.11.002
0010-0277/

Fig. 1. Example of a referential context. Photograph courtesy of Tamara Nicol Medina (Medina et al., 2011).

Trueswell, & Gleitman, 2011). If the learner's task in word learning actually required completely open-minded referent selection from this set of presented alternatives, surely language would be very difficult if not impossible to learn. However, paradoxically enough, Fig. 1 points to approaches for solving the very question it poses. After all, the infant in this picture is looking at the shoe beneath his walker. If parents tend to talk about what their children are attending to, the reference problem seems more tractable (Bruner, 1974/1975). Indeed, even outside observers of this snapshot of parent–child interaction guess quite often – and correctly – that the mother was uttering "shoe" at the moment the picture was taken. From this perspective, it seems hardly to matter how many objects, qualia, etc., are in reach of the visual scan – be it 10 or 1000 alternatives – what matters most for communication is the immediate "common ground", the focus of joint attention for the interlocutors (e.g., Grice, 1975, 1989; Lyons, 1999; see also Brown-Schmidt & Tanenhaus, 2008; Yoshida & Smith, 2008).

In the studies presented here, we aim to investigate the properties and behaviors present in parent-infant interactions that are informative for identifying the intended referent of child-directed speech. To do this, we examine parent-infant visual attention, gesture, and object manipulation as words are uttered under typical conversational circumstances in the home. Importantly, and as we describe further below (see Section 2.1), we take advantage of a particular property of our corpus: it includes an independent estimate of the referential transparency of each exchange. In particular, adult observers watched muted versions of these videos and guessed what words the parent was uttering, in a procedure known as the Human Simulation Paradigm (HSP, Gillette, Gleitman, Gleitman, & Lederer, 1999; Snedeker & Gleitman, 2003). This procedure provides us with an estimate of referential transparency as inferred from the extralinguistic cues present in each interaction – words that are easily guessed are assumed to have been uttered in more transparent circumstances than words that are more difficult to guess.

Our focus is on two interrelated questions. First, just how referentially ambiguous is the infant's (sampled) learning environment, operationalized as the HSP observers' ability to reconstruct the intended referent of words from whatever extralinguistic cues are present. Our second focus is on the role of the *temporal dynamics* of these interactions, i.e., how these extralinguistic cues intercalate in time with the word utterance itself. That is, following a venerable theme from David Hume (1748), we ask how precise temporally contiguous cues have to be for an observer to conclude that there is a cause-effect relation between input words and the nonlinguistic behavior of the speaker. Is the timing relation systematic and tight enough to support a learner's choice of referent among all those that are in principle available when scanning the passing scene?[1]

We are by no means the first to address these questions. The topic of joint attention and its explanatory role in language acquisition was introduced into the current experimental literature in a seminal paper by Bruner (1974/75) who suggested that joint attention and joint reference likely provided an important early mechanism for linguistic and social learning; parents might do much of the work of referent identification by talking about what children are attending to. These comments led to substantial observational research examining interactional cues to learning (Moore & Dunham, 1995, and papers therein), which revealed the social-attentive behaviors that arise in spontaneous parent–child interactions during object play, as recorded either in the lab or home (e.g., Harris, Jones, Brookes, & Grant, 1986; Tomasello & Farrar, 1986; Tomasello, Mannle, & Kruger, 1986; Tomasello & Todd, 1983). These now classic studies established that not all parental word utterances are created equal when it comes to their ability to predict child vocabulary growth and, by implication, to facilitate accurate referent identification. In particular, parents who engaged more in follow-in labeling – labeling what the child was currently attending to – had children whose vocabulary growth outpaced that of children who were exposed to proportionally more discrepant labeling situations, with the latter being negatively correlated with vocabulary growth (e.g., Tomasello & Farrar, 1986). This work suggests that, at least during controlled object play, referent identification is

[1] From the way we have just set our problem space, it should be clear that our primary interest in the present paper is the very beginnings of vocabulary learning, which relies much more on evidence from the co-present referent world. It is now well established that children make inferences about word meaning based not only on reference but on, e.g., collateral distributional and syntactic evidence (e.g., Chomsky, 1969; Landau & Gleitman, 1985; Lidz, Waxman, & Freedman, 2003; Naigles, 1990, inter alia). Yet, these linguistic resources cannot themselves be mobilized until a "seed" vocabulary, mainly of whole-object nominals are acquired by perceptual observation, and used to build distributional libraries and the syntactic structures of the exposure language (Gleitman, Cassidy, Nappa, Papafragou, & Trueswell, 2005). Reference finding via extralinguistic cue structure is only one evidentiary source for lexical learning but it is necessarily the earliest step, on which later accomplishments hinge.

best accomplished during episodes of joint attention[2] by parent and child.

Several subsequent laboratory experiments help to solidify our understanding both of the variety of extralinguistic cues to joint attention and their potential power in establishing word-referent pairings (e.g., Baldwin, 1991, 1993; Tomasello & Farrar, 1986). For example in her landmark study, Baldwin (1991) showed that infants (approximately 16–19 months) are sensitive to the attentional stance of their caregiver as assessed by his/her eyegaze, head-posture and voice direction. Infants showed signs of connecting a caregiver's utterance of a novel word to an object within the infant's current focus of attention if and only if the caregiver was also attending to that object; if the caregiver was attending elsewhere (i.e., discrepant labeling), infants avoided this word-referent mapping. Baldwin (1993) later found that older infants (19 months) could learn the caregiver's intended mapping even under discrepant labeling, when the speaker's visual target was occluded from the child's vantage point at the time of the speech act but then later revealed. Since that time, numerous experiments have documented the role of speaker and child attention in referent identification and word learning, corroborating and expanding on these early findings (e.g., Bloom, 2002; Jaswal, 2010; Nappa, Wessell, McEldoon, Gleitman, & Trueswell, 2009; Southgate, Chevallier, & Csibra, 2010; Woodward, 2003).

Informative as these experiments have been, the question remains how far laboratory settings, which radically reduce the referential choice set and script the conversational situation, can be linked to the dynamic and fluid circumstances of an infant's everyday life in which objects come and go in seconds and milliseconds, and words in sentences flow by at a rate of over 100 a minute. In response to this kind of concern, recent work has begun to examine the temporal dynamics of reference during unscripted object play. In many ways, this work returns to the earlier observational methods described above but now with an eye on the sensory and attentional mechanisms that support reference (e.g., Pereira, Smith, & Yu, 2014; Smith, Yu, & Pereira, 2011; Yoshida & Smith, 2008; Yu & Smith, 2012, 2013). The emphasis of this work has been on the child's perspective during these interactions, investigated via infant head-mounted cameras. Yoshida and Smith (2008) introduced and described this technology showing that the parent's and child's perspective on the same scenes differed systematically. After all, these infants and toddlers are very short and so may see the legs where parents see the tabletops. More importantly, many aspects and objects of the passing scene are not even in their purview. In these ways the infant has less (or distorted) information about what the mother is viewing and therefore talking about. Yet in other ways the infant is advantaged rather than disadvantaged in his perspective. He is receiving cues that go far beyond mere visual inspection – by moving their bodies and by grasping and holding objects of interest, infant-toddlers often bring only certain things into their visual focus. Single objects are then looming in front of them, occupying much of their visual field as the mother speaks of those things (Yoshida & Smith, 2008).

Using this method, Pereira et al. (2014) examined the sensory-attentional conditions that support word learning during dynamic object play between parent and infant in the laboratory. The parent was first taught the names of three novel objects. Rather than providing a script, the experimenters asked parents to talk "naturally" with their child about the new objects as they played with them on a tabletop, while the child's view was recorded. The immediate visual environment, from the child's eye view, was examined as it co-occurred (or did not co-occur) with the introduction of object labels. Afterward, children were tested on their understanding of these labels: They were asked to pick out the novel object based on its name (e.g., "Show me the groodle"). Accurate referent selection was predicted by certain sensory conditions during earlier parent naming: learned words tended to be the ones that mother had uttered when the object was more centrally located and looming large in the child's view, approximately 3–4 s before and after the naming event, whereas unlearned words tended not to have these characteristics. An additional detail of great potential explanatory importance: effective learning instances were also characterized by sustained child attention immediately following the naming event, suggesting that sustained child examination of the object ("sticky attention" is the phrase coined to describe this state) is helpful for consolidating the word-referent pairing in memory (Lawson & Ruff, 2004; Vlach & Sandhofer, 2014).

Thus Pereira et al. (2014) offer a possible approach to how referent-matching can happen in a complex world. The learner is selective in attention and thus can avoid being bombarded by countless distracting things, happenings, qualities, and relations. The learner has some implicit criteria (perhaps something like "single looming object visuo-centrally in view") that heavily constrain the choice between relevant and irrelevant potential referents. This suggests that at least some natural early naming events are, for practical purposes, not ambiguous, and it is these moments that move learning forward. These findings are consistent with earlier laboratory work examining the importance of temporal contiguity between linguistic input and object attention (e.g., Gogate, Bahrick, & Watson, 2000; Hollich et al., 2000; Jesse & Johnson, 2008).

The studies we present here are very much in line with Pereira et al., 2014, except that we ask what the dynamics of referent identification are like during naturally occurring utterances in the home, when parents utter common content nouns to their children. With very few exceptions, past observational work on reference identification has examined facilitating learning conditions under a single restricted circumstance: object play on the floor or on a tabletop (e.g., Harris et al., 1986; Pereira et al., 2014; Tomasello & Farrar, 1986; Tomasello & Todd, 1983; Tomasello et al., 1986) – potentially limiting the generality of the observations. Past research has also typically focused on those parental utterances that are about co-present objects of experimental interest – i.e., only the labeling of objects that were provided by the experimenters. Despite their clear usefulness, these approaches did not try to assess how frequent or prevalent, in the child's everyday world, are these referential moments of clarity.

Indeed, there is other evidence to suggest that the extralinguistic visuo-social context of caregivers' utterances offers only rare moments of word-referent clarity, at least for the child learning her first words. This evidence comes from studies employing the HSP (Human Simulation Paradigm, Gillette et al., 1999) that, as mentioned above, asks adult observers to watch muted video examples of parents spontaneously uttering a word of interest to their children (e.g., "Give me your foot"). In these "vignettes", a

[2] Akhtar and Gernsbacher (2007) have argued that although joint attention can play a role in word learning, it is not actually a requirement, especially if the definition of joint attention requires not only that focus of attention (as defined, e.g., by Baldwin, 1995, and similarly by Tomasello, 1995) between listener and speaker be shared, but that both interlocutors be aware of this. Akhtar and Gernsbacher review evidence that word learning "can occur without joint attention in typical development, in autistic development and in Williams Syndrome and joint attention can occur without commensurate word learning in Down Syndrome". Similarly, Yu and Smith (2013) argue that merely overlapping attention, which they called "coordinated attention", is enough to foster learning. Following this idea, when we refer to joint attention in the remainder of this paper, we mean that parent and child are looking at the same object; if we mean something more than that, such as the requirement of awareness or intention to refer, we will indicate this explicitly. Moreover, we acknowledge that learning, even by babies, can occur incidentally and in response to diffuse situational factors. The joint attention condition is facilitative – heavily so, especially early in the learning process – but not invariably required.

beep is played at the exact moment the caregiver uttered the word, with the observer then guessing what the mother must have said. Videos are muted so as to simulate in the adult observer what it is like for a child learning her first words; the child does not know the meaning of the other words in the utterance and thus cannot use these to determine meaning. The logic then is that the more informative the interaction between parent and child was, the more accurately the HSP participants should guess what word the parent had actually uttered at the time of the beep. Of particular interest for the present work are the findings of two recent HSP studies: Medina et al. (2011, Exp. 1) and Cartmill et al. (2013). These researchers video-recorded examples of parents interacting with their infants (12–15 mo. olds for Medina et al.; 14–18 mo. olds for Cartmill et al.). Unlike most observational studies that focus on object play, these researchers recorded interactions in the home during unscripted everyday events, including meal, bath, and play time, etc., where object play may or may not have been happening. Rather than focusing only on caregivers' utterances that mentioned co-present objects of interest, the researchers used randomly selected examples of the caregivers uttering common content words. Thus, at times, a referent was not even present, e.g., when a mother talks about the bird they saw at the zoo. The result is a more representative sample of the contexts in which infants encounter acoustically salient content words in speech directed toward them.

Both Medina et al. (2011) and Cartmill et al. (2013) found that events of word-referent clarity are rare in these everyday circumstances. Of the 144 vignette examples of parents uttering common nouns, Medina et al. found that only 20 (14%) were "highly informative" referential acts, defined as vignettes for which 50% or more of the HSP observers correctly guessed the target word. The vast majority of vignettes were instead "low informative" (guessed correctly by less than 33% of HSP observers). And, although not reported in the paper, an analysis of the dataset shows that, across all noun vignettes, average HSP accuracy was 17% correct. Cartmill et al. (2013, whose data we analyze further below in Study 1) report a slightly higher average HSP accuracy of 22%, when sampling 10 concrete noun utterances produced by each of 50 SES-stratified families. As we report below, only a small percentage of vignettes were highly informative. Notably, the HSP averages from Medina et al. and Cartmill et al. are slightly lower than the 28% accuracy found in the first reported HSP study from Gillette et al. (1999, p. 147, Fig. 2, noun data, trial 1). But, Gillette et al. used examples of common content nouns uttered only during object play, suggesting that this situation elevates referential clarity. Interestingly, a recent HSP study by Yurovsky, Smith, and Yu (2013) reports a noun accuracy of 58% correct, but this study used videos of object play *and* only sampled utterances for which the caregiver labeled a co-present object of interest – suggesting that, as one might suspect, labeling a co-present object during object play can elevate referential transparency.

Thus, taken together, HSP results suggest that common content nouns, when uttered by parents to their infants under everyday circumstances, offer moments of referential transparency only on rare occasions, with the results of Medina et al. suggesting it is about 1 in 6 times. As reported below, the present work re-affirms this rarity, even when limiting the utterances to words not found in the child's productive vocabulary. But more centrally, we ask what visual-attentive behaviors, and their timing, characterize referential clarity during these naturally occurring utterances. Study 1 re-analyzes a subset of the HSP data collected by Cartmill et al. and then codes these same videos for known cues to referential intent, including referent presence, parent and child attention, and gesture. By relating these codes to HSP accuracy, we can identify what behaviors, relative to word onset, promote accurate referent identification. Study 2 reports a new HSP experiment designed to examine in detail how important the relative timing is between word occurrence and these extra-linguistic behaviors. As we shall see, our results from these everyday circumstances will, in many ways, be in line with observations made under controlled laboratory settings, including recent timing results from 1st person cameras (Pereira et al., 2014); this fortunate alignment of results occurs even though our own work comes from stimuli recorded from a 3rd person view – an issue to which we return in the General Discussion.

2. Study 1: Timing of cues to referential intent in parent–child interactions

We begin with a coding analysis of an existing HSP video corpus, developed by Cartmill et al. (2013) and introduced earlier. The entire corpus consists of 560 40-s video clips, each containing an example of a caregiver uttering a common content noun to his or her 14- or 18-mo old. These conversational interactions are from 56 English-speaking SES-stratified families (10 vignettes each), collected as part of a larger longitudinal study of language development (see Goldin-Meadow et al., 2014). Each vignette has associated with it the responses (as collected by Cartmill et al.) from approximately 15 adult HSP observers who guessed what the caregiver was saying from muted versions of the videos. Here we report a new set of analyses on a subset of these vignettes.[3] As we were interested in potential word-learning moments, we restricted our analyses to examples of parents uttering nouns that were unattested in the child's own speech at the time of the home visit (see details below). Moreover, because we wished to document the visuo-social behaviors leading up to the word's occurrence in each video, we included only those videos for which 14 s had elapsed before the first utterance of the target word.

The result is an analysis of a corpus containing 351 40-s vignettes, each of a caregiver uttering a common concrete content noun to his or her infant under everyday circumstances. We report two findings. First, we report a re-analysis of the HSP data collected by Cartmill et al. but now focusing on these 351 vignettes, so as to determine with what frequency highly informative referential acts occur in this sample (note that Cartmill et al. did not categorize vignettes into "high" or "low informative"). Second, we report results from two trained coders, who coded on a second-by-second basis potentially important aspects of each video – referent presence; parent- and child-attention to the referent; parent- and child-attention to other objects; parent gesture to and/or presentation of the referent; and parent–child joint attention. Here we report which of these aspects, and their relative timing properties, reliably characterize highly informative, referentially transparent acts as operationalized by HSP accuracy.

2.1. HSP video corpus

Details of the corpus, including video selection criteria and how the adult HSP data were collected, are reported in Cartmill et al. (2013). For clarity, some of this information is repeated here (but see Cartmill et al. for additional details). In brief, the videos come from 56 families participating in a longitudinal study of language development (Goldin-Meadow et al., 2014). All children were typically developing (30 males, 26 females) and were being raised as monolingual English speakers. As part of the longitudinal study, families were visited in their homes every 4 months from child

[3] Cartmill et al. report HSP data from only 50 of the 56 families; they excluded 6 families because they did not have measures of vocabulary outcomes necessary for their analyses. As these measures are not part of the current study, we re-included the HSP data from these families, and the corresponding videos, bringing our family total back up to 56.

age 14 to 58 months, and were video recorded for 90 min at each visit. During visits, families engaged in their normal daily activities, ranging from book reading and puzzle play to meals, bathing, and doing household chores.

The Cartmill et al. HSP corpus consists of 560 forty-second videos ("vignettes"), 10 from each of the 56 families – 6.2 h in total. These vignettes came exclusively from the 14- and 18-mo-old visits. Each vignette was an example of a parent uttering one of the 41 most common concrete nouns in the entire sample from these visits, usually produced within a sentence context (e.g., Can you give me the book?). Vignettes were aligned so that 30 s into the video, the parent uttered the target word (at which point a beep was inserted). If the parent uttered the target word more than once during the 40 s vignette, each instance of the target word was marked by a beep. To select vignettes, Cartmill et al. (2013) ranked concrete nouns uttered to these children at 14–26 months by frequency, and randomly chose a single example of the 10 highest-ranked words each parent produced at child age 14–18 months. Because highest-ranked nouns varied across parents, the final test corpus contained 41 different nouns.

Each vignette has associated with it the responses of approximately 14 to 18 native English-speaking adults who participated in the HSP study reported in Cartmill et al. (2013). In that study, a total of 218 participants (145 female) were randomly assigned to one of 15 experimental lists, each consisting of 56 vignettes (including both targets and filler vignettes, which were examples of abstract nouns and verbs). Participants were undergraduates enrolled either at the University of Pennsylvania or La Salle University in Philadelphia. After viewing a vignette, participants guessed the "mystery" word for that vignette before viewing the next vignette. Participants were tested individually or in groups, ranging from one to six people. Video was projected on a wall or screen and participants recorded their guesses on paper. Cartmill et al. (2013) scored a participant's guess as correct if it was identical to the target word. Abbreviations and plurals were also counted as correct (e.g., phone or phones for telephone), but words that altered the meaning of the root word were not (see Cartmill et al. for further details).

Our analyses included 351 of the 560 vignettes, selected on the basis of two criteria. First, we selected only vignettes for which the word was not attested in the child's own speech at the time of the recording, as determined by parent responses to the MacArthur Communicative Development Inventory (CDI) and by the child's own productions during the 14- and 18-month home visits. This criterion reduces the possibility that the child's response to a familiar word (e.g., grasping a ball after hearing "pick up the ball") would offer an unfair clue to the HSP participants as to the intended referent (for discussion and analysis of findings partitioned according to this distinction, see Cartmill et al., 2013). Second, in order to be able to examine behavior leading up to a word's occurrence, we included only vignettes that had at least 14 s of video prior to the first word occurrence (i.e., 14 s of silence before first beep). These criteria resulted in 35 word types in total (see Table 1).

Table 1
Target words.

Study	Words
Study 1	ball, bear, bed, bird, block, book, bowl, button, car, cat, chair, cheese, cookie, cup, dog, door, eye, face, fish, foot, hair, hand, head, juice, kiss, milk, mouth, nose, orange, phone, pig, shirt, shoe, step, water
Study 2	ball, bear, block, book, car, cat, cheese, cookie, dog, door, duck, eye, hair, hand, kiss, mouth, nose, phone, shoe, water

2.2. Coding the corpus for extralinguistic cues to reference

Two trained coders viewed the muted vignettes using ELAN (http://tla.mpi.nl/tools/tla-tools/elan/; Lausberg & Sloetjes, 2009).[4] At the time of coding, coders were blind to the HSP accuracy associated with each video. Each video was coded for the following.

(1) *Presence of Target Referent:* Target referents were coded as present when they were visible on the screen and could be easily and accurately identified. In cases where the referent was partially obscured, blurry, or difficult to recognize, object presence was coded as *maybe* and treated as present during analysis. In most cases of this sort, the referent was within the child's ability to see, with the camera-work responsible for the blur or bad angle (cf. Yurovsky et al., 2013).

(2) *Parent Attention to Target Referent and Other Objects:* Parent attention was coded as present when a parent attended to an object through (1) overt visual focus on the object (when the eyes could be seen), (2) head or body orientation toward the object (when the eyes could not be seen), (3) physical interaction with the object, or (4) deictic gesture toward the object. In case of conflict between visual focus and body orientation (e.g., looking at a ball while the body is oriented toward a toy train), the object that was given overt visual focus was considered to be the target of parent attention. In the absence of overt visual focus, physical interaction with an object (e.g., holding, touching, shaking or playing with the object) was taken to reflect attention. Physical interaction could not be incidental or passive contact with the object (e.g., sitting on a chair). In terms of gesture, only deictic behaviors were coded (e.g., pointing toward an object or holding the object up). We coded attention to off-screen objects if and only if they later became visible without a break in attention (e.g., if a parent looked off camera at a dog who then entered the scene, attention was coded from the start of the look rather than the moment the dog appeared within the frame). The target of attention was always assumed to be a whole object (e.g., a toy bear) unless focus on a specific component was overtly signaled through close visual inspection or physical manipulation (e.g., pointing to or wiggling the ear of the toy bear). People were not considered possible referents of attention unless a specific body part or clothing item was highlighted. Attention was coded as continuously directed toward a single referent until a break in attention of 2 s or more was observed. Since attention to the other conversational participant was not coded, periods of time when the parent was attending to the child were coded as no parental attention.

(3) *Child Attention to Target Referent and Other Objects:* Child attention was coded using the same criteria used for Parent Attention.

(4) *Parent Gesture/Presentation of Target Referent:* Gesture/presentation of target referent was a subset of Parent Attention to Target Referent, and was defined as any parent action or gesture used with the intention or effect of calling another person's attention to the target referent. Gesture/presentation was coded as present from the onset of a gesture toward or action involving the object until the gesture was retracted or contact with the object was broken. Again, only deictic gestures were coded. Presentation of objects included actions or motions directed toward the target referent that

[4] Sixteen vignettes were excluded due to stimulus preparation errors; in these, the beep had been incorrectly positioned in the video.

might function to highlight the referent in some way (e.g., offering, reaching, touching, grabbing, shaking, and holding out).

Each coder completed a subset of the vignettes (112 were double-coded). Codings were binary scores (1 = presence) for every second of each video. Mean proportion of agreement for each coded cue was 0.89 (range 0.82–0.99) with inter-rater reliability 'good' to 'near perfect', per Landis and Koch's (1977) descriptive division (mean Kappa 0.73; range 0.63–0.83). In cases of disagreement, the code used for analyses was 'present'. For all subsequent analyses, codes were treated as binary (present = 1; absent = 0).

Note that in order to mark attention to the target referent, our two coders had to know the target word. To address the coding bias that this procedure could have introduced, we ran a separate experiment in which 12 participants coded child target attention for 20 vignettes (10 word types, 2 each), with participants told (n = 6) or not told (n = 6) the target words. Kappa-agreement scores between individual subjects and our two coders ranged from .77 to .87 (M = .83, SD = .03), corresponding to Landis & Koch's 'good' to 'very good' agreement for each subject. Also, these Kappa scores did not differ between the "told" and "not told" subject groups, as assessed by unpaired t-tests, suggesting that both groups were in similar agreement with our two coders.

2.3. Results and discussion

The results are divided into two sections. First we report our re-analysis of the HSP responses from Cartmill et al. (2013), focusing only on examples of parents uttering words not attested in the child's productive vocabulary. Second, we report new analyses of the coded extra-linguistic cues to reference and their timing, and we relate HSP response accuracy to coded cues in order to determine which behaviors and scene properties support word-referent clarity.

2.3.1. HSP responses: Rarity of referentially transparent nominal events

As reported elsewhere with this dataset (Cartmill et al., 2013) and similar datasets (e.g., Medina et al., 2011), we found that, in our subset of Cartmill et al. (2013) vignettes, most contextual environments for word utterances were relatively uninformative when it comes to identifying the speaker's intended meaning: the average proportion of correct HSP guesses (proportion of token guesses) per concrete noun was just 0.18 (Table 2); that is, only 18% of all HSP responses were correct when averaging across all responses.

Following Medina et al. (2011), we wanted to examine for this data set just how often a vignette would be considered "highly informative" for correct referent identification. We therefore split vignettes into three categories based on HSP accuracy: Highly Informative (HI) vignettes (those with HSP accuracy greater than or equal to 50%); Low Informative (LI) vignettes (accuracy less than or equal to 10%) and Middle Informative (MI) vignettes (accuracy greater than 10% and less than 50%). This split was not reported in Cartmill et al. (2013). Our question was whether we would replicate in this new sample Medina et al.'s observation that HI vignettes are rare (occurring about 1 in 6 times), except here using only words not attested in each child's productive vocabulary. Indeed, as seen in Table 2, row 1, HI vignettes make up 14% (50 out of 351) of the vignettes, with the majority defined as low informative (LI, 58%, 202 out of 351).

Table 2 also presents an analysis of the mean number of HSP response *types* per vignette (row 2), along with the minimum (row 3) and maximum number (row 4). For this analysis we calculated how many different types of responses were offered by HSP participants per vignette; e.g., if 7 participants responded "shoe", 5 responded "dog" and 3 responded "cat", the number of types for that vignette would be 3 ("shoe", "dog", "cat").[5] (see Appendix A for examples of actual HSP responses for HI, MI and LI vignettes.) This particular measure gives us an estimate of the number of different referent types that might come to mind for a learner from each caregiver utterance, given the extra-linguistic context. As can be seen in Table 2, the average number of response types is 7.6 (row 2), with a range of 1 to 14 (rows 3 and 4). To the extent that this measure can be taken as an estimate of the types of word meanings that come to mind for the child in the video, a picture emerges suggesting that most situations in which common concrete nouns are uttered are ones that offer many referential alternatives; very rarely are a small number of alternatives available (e.g., the average number of response types for HI vignettes is 4.3, see row 2). Note also that some rare situations can be misleading to the learner: in particular, some rare LI vignettes have a low number of response types (i.e., a minimum of 1, see row 3); it is simply that these referent guesses are incorrect, though consistent with one another. It is also worth noting that what is happening in the vignettes is pertinent to what is actually being talked about––100% of the 149 MI and HI vignettes had at least one HSP participant guess the correct word, although only 16% of the LI vignettes (33 of 202) were ever guessed correctly (see row 6 of Table 2).

Table 2
Human Simulation Paradigm (HSP) responses.

Measures	Type of vignette			Overall
	LI	MI	HI	
1. Number of vignettes (n)	202	99	50	351
2. Average # of response types (out of 15)	8.34	7.77	4.30	7.60
3. Minimum # of response types (out of 15)	1	3	1	1
4. Maximum # of response types (out of 15)	13	14	8	14
5. Proportion correct responses (HSP accuracy)	0.01	0.26	0.72	0.18
6. Vignettes at least one correct response	33	99	50	182

Note: LI = Low Informative vignettes; MI = Middle Informative; HI = High Informative.

In sum, our analysis of the Cartmill et al. (2013) HSP responses suggest that most utterances containing concrete nouns that parents produce in the quotidian environments of the home are relatively uninformative from the point of view of word learning; a small minority of utterances occur in an environmental context that permits most people to guess what the parent must have meant. As we have argued elsewhere (Cartmill et al., 2013; Medina et al., 2011), these specially informative instances are likely to create the kinds of potent learning opportunities that drive children's early vocabulary growth (see Section 4). As such, it becomes interesting to ask which aspects of the environmental context make these learning instances, and not others, highly informative. We turn to this issue next.

2.3.2. Properties of the environmental context

Vignettes were coded on a second-by-second basis for the following properties: (1) Presence of Target Referent; (2) Parent Attention to Target Referent and Other Objects; (3) Child Attention to Target Referent and Other Objects; (4) Parent Gesture toward and/or Presentation of Target Referent. We asked how well these

[5] This measure was called the spread of HSP responses in Gillette et al. (1999) whose findings of spread for common nouns were in this same numerical range. Note also that in the present study there were an uneven number of HSP subjects contributing to each vignette. The average was 14.5 subjects (range 12–21). We therefore normalized all results to be a maximum of 15 subjects by calculating the type-token ratio (number of Types divided by N) and multiplying by 15. Because the number of subjects (N) did not vary much between items (for 307 of the items, N = 13, 14, 15 or 16), the results are essentially identical to using raw counts per vignette.

codes predict HSP accuracy scores, i.e., what characteristics make the intended referent obvious to naïve observers, and whether the characteristics appear at certain moments in the vignette relative to when the word was uttered. We first considered the contributions of each property separately. Tables 3A and B and Fig. 2 summarize the key findings and will be used as touchstones for succeeding discussions. Table 3A presents coarse grain averages of each code, indicating the presence (yes/no) and average duration of a property during a large time window of 20 s before and 10 s after target word onset. These measures are split by HSP accuracy level of vignette (LI, MI, HI); accuracy level divisions are somewhat arbitrary and are merely presented to illustrate the numerical relation between coded scene properties and referential informativity of the vignette. Table 3B presents corresponding statistical analyses: separate simple linear regressions (10 in total) in which the coded measure of each vignette was used to predict its HSP accuracy score. HSP accuracy for each vignette (a proportion from 0.0 to 1.0) was first transformed to an empirical-logit (elogit) value, as were all proportion of time measures. Fig. 2 presents more fine-grained temporal analyses of each code. In each panel, the proportion of vignettes with a coding property present has been plotted on a second-by-second basis relative to the onset of the target word,[6] split by vignette type (LI, MI and HI). Reliability of these observations was determined by treating (elogit) HSP accuracy as a continuous variable and testing how well it was predicted by the probability of each annotated code on a second-by-second basis. Shaded areas in Fig. 2 indicate those seconds for which the correlation was reliable, after Bonferroni-adjusting for multiple tests (i.e., all p's < 0.00027). See Appendix B for details of all tests.

Table 3A
Average coarse-grain properties of the environmental context in which a target word was uttered; words are classified according to their HSP accuracy, and context was coded starting 20 s before and continuing through10 s after word onset.

Dependent variable	Type of vignette			Overall
	LI	MI	HI	
Target referent is present (no = 0, yes = 1)	0.69	0.93	1.00	0.80
Proportion of time target referent is present	0.54	0.67	0.70	0.60
Parent attended to target referent (no = 0, yes = 1)	0.10	0.23	0.56	0.21
Proportion of time parent attended to target referent	0.02	0.03	0.07	0.03
Child attended to target referent (no = 0, yes = 1)	0.23	0.49	0.82	0.39
Proportion of time child attended to target referent	0.05	0.14	0.22	0.10
Parent gestures to or presents target referent (no = 0, yes = 1)	0.23	0.51	0.74	0.38
Proportion of time parent gestures to or presents target referent	0.07	0.21	0.26	0.14
Co-incident attention to target referent (no = 0, yes = 1)	0.11	0.28	0.58	0.23
Proportion of time co-incident attention	0.02	0.07	0.11	0.05

Note: LI = Low Informative vignettes; MI = Middle Informative; HI = High Informative.

Table 3B
Results of separate linear regressions. Coarse-grain environmental property of each vignette was used to predict the vignette's elogit of HSP accuracy (−20 to +10 s from word onset). All correlations are statistically significant, even after Bonferroni-adjusting for 10 tests (i.e., all p's < 0.005).

Predictor of HSP	Est.	Std. error	t	p-value
Target referent is present (no = 0, yes = 1)	1.44	0.21	6.98	<.0001
Elogit of time parent attended to target referent	0.11	0.03	3.93	.0001
Parent attended to target referent (no = 0, yes = 1)	1.58	0.16	9.96	<.0001
Elogit time parent attended to ref	0.44	0.05	9.34	<.0001
Child attended to target referent (no = 0, yes = 1)	1.40	0.16	8.58	<.0001
Elogit of time child attended to target referent	0.34	0.04	8.24	<.0001
Parent gestures to or presents target referent (no = 0, yes = 1)	1.68	0.20	8.48	<.0001
Elogit of time parent gestures to or presents target referent	0.66	0.08	7.82	<.0001
Co-incident attention to target referent (no = 0, yes = 1)	1.58	0.19	8.20	<.0001
Proportion of time co-incident attention	0.49	0.07	7.54	<.0001

2.3.2.1. Presence of target referent. It will come as no surprise that an object fitting the description of the uttered concrete noun appeared in the majority (80%) of vignettes at least at some point during the 20 s prior to and 10 s after word onset (see first row of Table 3A, Referent Presence Overall). As shown in this same row, LI vignettes were less likely than HI vignettes to have the target referent present (69% vs. 100%), which resulted in a reliable effect of referent presence on HSP accuracy (Effect of Referent Presence on HSP, first row of Table 3B). In addition, the proportion of time the target referent was present during this large 30 s time period (second row of Table 3A) was reliably related to HSP accuracy in the expected direction (second row of Table 3B).

More interestingly, fine-grained temporal properties characterized the presence of the target referent for HI vignettes. As seen in Fig. 2A many HI vignettes are characterized by the sudden appearance of the referent just prior to, or at, word onset. Initially, LI, MI and HI vignettes all have the referent present with about a 0.5 to 0.6 probability. Only HI vignettes, however, show a sharp rise in referent presence just before the parent uttered the target word, peaking nearly at a 1.0 probability one second after word onset, with target referent presence maintaining a high probability for several seconds after word utterance. This pattern resulted in a reliable effect on HSP accuracy scores of referent presence starting 2 s before word onset and continuing through the rest of the vignette (as indicated by the shaded area in Fig. 2A). What this suggests is that the mere presence of the target referent during a large portion of a parent–child interaction is not predictive of a referentially transparent act; rather it is the sudden, and sustained, appearance of the object just prior to word utterance that is predictive. Although the continuous presence of the referent is not predictive, vignettes with continued presence of the referent may have other properties, for example, attentional cues discussed next, that could make them highly informative.

2.3.2.2. Parent attention to target referent and other objects. As shown in the third and fourth rows of Table 3A, parents showed an increased probability of attending to the Target referent and spent more time attending to this referent in HI as compared to LI vignettes during the broad window of −20 to +10 s relative to word onset, resulting in reliable effects on HSP accuracy of both measures (see corresponding rows of Table 3B).

Here once again though, presence of these properties is closely time-locked with the onset of the relevant word. As shown in Fig. 2B, a large subset of HI vignettes involve a sudden shift in parent attention to the target referent just prior to word onset, which

[6] For vignettes with more than one beep (i.e., when a parent uttered the target word more than once during the vignette), timing was calculated relative to the first word occurrence. It is conceivable that one of the additional beeps heard by our HSP subjects helped or hindered correct identification, but this would only add noise to our analysis by misclassifying an otherwise HI vignette as LI or vice versa. Consistent with this, a separate analysis on only those vignettes with a single beep (n = 202) generated timing patterns very similar to those in Fig. 2, with similar statistical patterns but with less power. For these reasons, we present all the data here rather than the subset of one-beep vignettes.

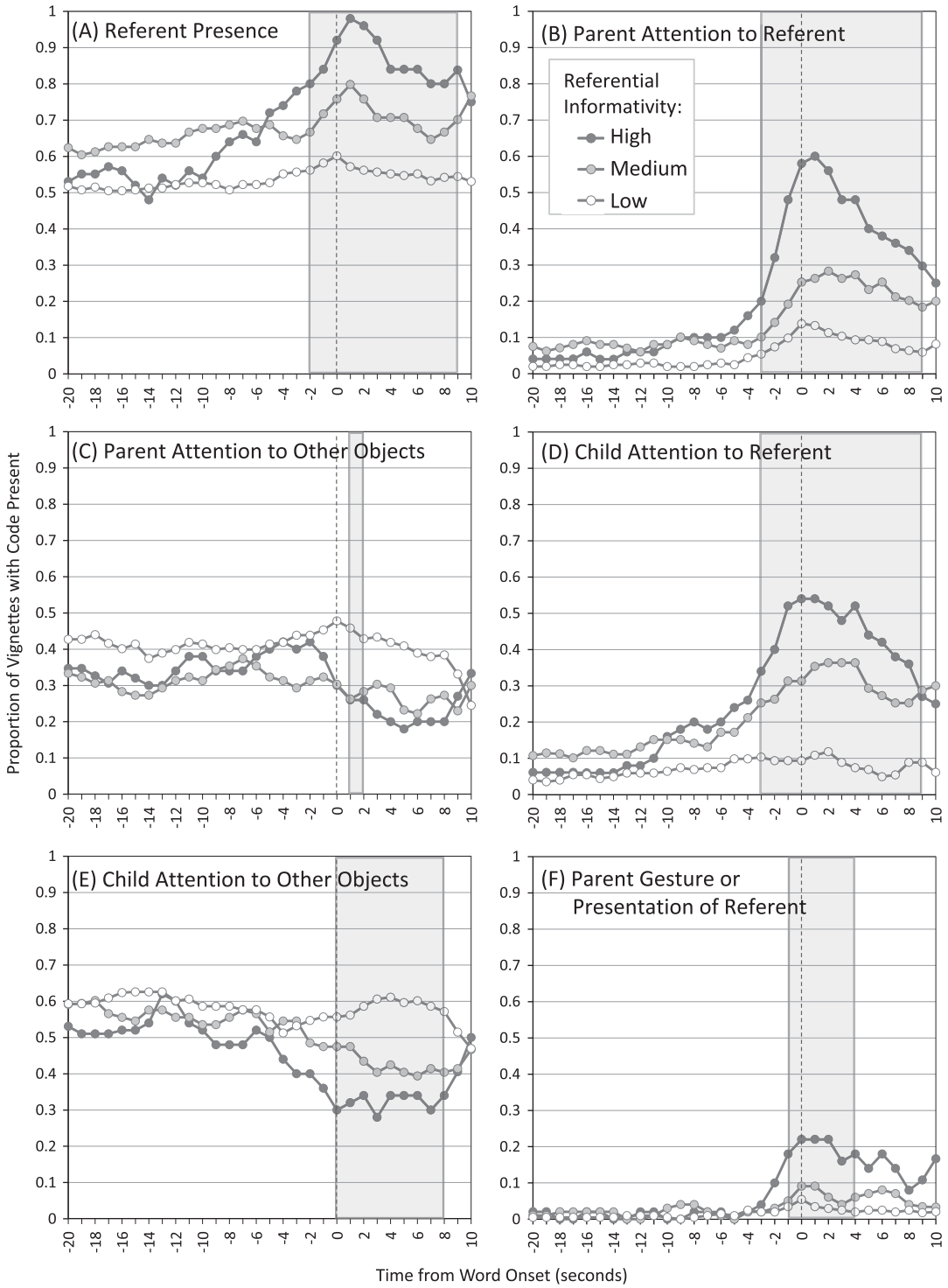

Fig. 2. Proportion of vignettes with social-attentive cues coded as *present* (*y*-axis) in relation to word onset (dotted line on the *x*-axis). Shaded areas indicate periods during which presence/absence of the coded cue reliably predicted HSP accuracy scores for each vignette as based on separate Generalized Models (all *p*'s < 0.00027). At least 14 s of the video stream occurred before the first utterance of the word (hence: before the first beep), and it is this word occurrence that always aligns to 0 s (see text).

is sustained after word onset. This pattern resulted in parent attention to target referent reliably predicting HSP scores from −3 s to +10 s from word onset. It is also revealing that parent attention to non-target referents trends (though not reliably) in the reverse direction (Fig. 2C). These findings are consistent with laboratory and other observational work suggesting an important role for parental attention in highlighting a referent, but here we show that the timing of these attentional cues plays an important role, and that these cues are present in only a small subset of (highly informative) interactions.

2.3.2.3. Child attention to target referent and other objects. As shown in the fifth and sixth rows of Table 3A, children were more likely to attend to the Target referent, and spent more time attending to this referent, in HI as compared to LI vignettes, resulting in reliable effects on HSP accuracy of both measures (see corresponding rows

of Table 3B). Here too these properties are closely time-locked with word onset. As shown in Fig. 2D, a large subset of HI vignettes involve a sudden shift in child attention to the intended referent just prior to word onset, continuing after word onset. This pattern resulted in child attention to the target referent reliably predicting increased HSP scores from −3 s to +10 s from word onset. It is noteworthy that child attention to non-target referents showed the opposite pattern, reliably for 0 s to +8 s from word onset (Fig. 2E).

Thus we see signs of follow-in labeling of the kind created in the laboratory by Baldwin (1991) and others. Parents will sometimes talk about what children are already attending to. When this condition is satisfied, and if the timing of the parental utterance is tightly time-locked to this turn of child attention, it appears to be easier for observers to recover the speaker's intended referent. However, we also find that this condition is rarely satisfied in our video slices of real life, i.e., given the paucity of high informative learning instances in the sampled input; the contribution of this factor to word learning has to be evaluated in light of this rarity of occurrence.

2.3.2.4. Parent gesture/presentation of referent. Parental gesture to and/or presentation of the target referent also can be observed in many HSP videos; this factor of close engagement with the referent is also correlated with high informativity. As shown in the seventh row of Table 3A, over half (56%) of all HI vignettes included the parent gesturing at or presenting the target referent during the broad window of −20 to +10 s relative to word onset, whereas LI vignettes had parent gesture/presentation of the target referent on just 10% of trials. Similar patterns can be seen in the proportion of time parents spend gesturing/presenting the referent (row 8 of Table 3A). Table 3B indicates that both of these measures are reliably related to HSP accuracy in the expected direction.

Again it is no surprise that close engagement with an entity serves as a cue that it might be what's being talked about. What is potentially more noteworthy, however, is that the temporal characteristics of gesture/presentation were precisely time-locked to word utterance. This effect is shown in the time-course plot of Fig. 2F. HI vignettes were associated with an increased probability of gesture/presentation near word onset, resulting in this coding measure predicting HSP scores from −1 s to +4 s from word onset. Finally, it's important to keep in mind that such cues are rare, usually only in highly informative interactions (see Section 4).

2.3.2.5. Referent presence vs. attention to referent. One might wonder if increases in attention to the target referent observed in panels B through F in Fig. 2 are a product of the object suddenly appearing (panel A), since Parent/Child Attention seems to go hand-in-hand with increases in Referent Presence. This, however, is not the case: Even when we restrict analyses to videos where the referent was present over 95% of the time ($n = 143$), we find attentional patterns very similar to what we see in Fig. 2 (see Appendix C, Fig. C.1). Thus, sudden appearance of an object, and sudden attention to an object, are independent positive cues to referential intent.

2.3.2.6. Joint attention and co-incident attention. Thus far, the analyses of parent and child attention have assessed only their independent contributions. Now we consider how parent and child attention routinely overlap in vignettes as a function of vignette informativity. For instance, it is possible that the same HI vignettes that show increases in child attention also show increases in parent attention, the condition of mutual-attention that we term *co-incident attention*, when both interlocutors' attention is fixed on the same entity simultaneously. We purposely avoid the term joint attention here because most researchers use the term joint attention to include both co-incident attention and what might be called staggered attention, in which the parent and child go through bouts of each separately attending to the same referent.

As shown in the final two rows of Table 3A and B, HI vignettes, relative to LI vignettes, are characterized by an increased probability of co-incident attention and increased time in co-incident attention. As shown in Fig. 3, HI vignettes are characterized by a higher probability of co-incident attention near word onset, resulting in a reliable effect on HSP accuracy of co-incident attention from −2 s to +7 s from word onset. The numerically lower values here compared with those for parent-attention to referent (Fig. 2B) and child attention to referent (Fig. 2D) suggest that co-incident attention also plays a role in referential informativity. Moreover, like the other measures, these events are all restricted to the rarer HI (and some MI) events.

2.3.2.7. Cue combination: As estimated by multiple regression. The previous two sections provided targeted examinations of cue combination and cue interaction – specifically examining what our coded behaviors looked like when the referent is present throughout (Section 2.3.2.5) and the time-course and informativity of over-lapping co-incident attention (Section 2.3.2.6). A more formal analysis is possible by examining the multiple simultaneous contributions of various coded cues to predicting HSP accuracy, done within a multiple regression. That is, rather than doing separate simple linear regression models in which each code is used to predict HSP on a second-by-second basis, here we report the results of multiple linear regressions (also on a second-by-second-basis) in which we use our coded events to simultaneously predict HSP accuracy. The advantage of such a model is that it can provide a picture of which of the coded events better predict HSP accuracy and when.

For simplicity, we have chosen a model which we believe most logically represents the structure of our data, namely, a model in which HSP accuracy scores are simultaneously predicted by: (1) Referent Presence (no/yes); (2) Parent Gesture/Presentation of Referent (no/yes), (3) Parent Attention to Referent (no/yes); (4) Child Attention to Referent (no/yes); and (5) the interaction between Parent and Child Attention to Referent (a.k.a., Co-incident Attention) (no/yes). This model was applied on a second-by-second basis (−20 s to +10 s word onset). A total of 16 of the 31 models (from −5 s to +10 s word onset) were reliably different from the corresponding null model in chi-square tests after Bonferroni correcting for the number of tests (31).

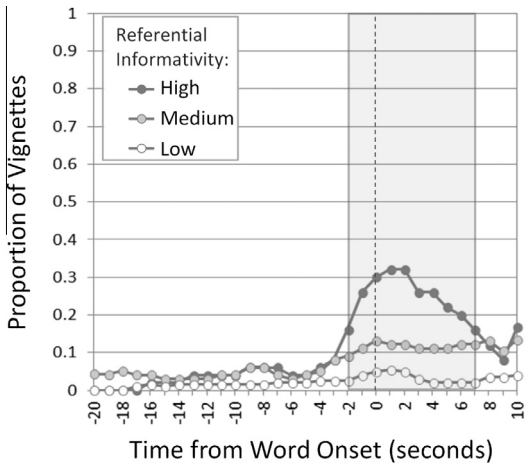

Fig. 3. Proportion of co-incident attention of parent and child, split by HI, MI and LI vignettes.

The details of these models appear in Appendix D. There one can inspect when each predictor in question was significant in the multiple regression that included all other predictors ($p < 0.01$). These patterns are very similar to the reliability of the simple regressions (which were reported earlier as the shaded areas in Fig. 2), with the only exception being that parent gesture/presentation is no longer a significant contributor at any time point. The implication here is that no single coded behavior drives the perception of referential transparency (HSP accuracy scores). Rather, several factors, particularly Referent Presence and Parent and Child Attention to the Referent near word onset and thereafter, are simultaneous contributors. The fact that Parent Gesture/Presentation does not contribute to the multiple regression analyses may not be too surprising since this measure is a logical subset of Parent Attention to Referent. Note the opposite could have occurred in principle: that Parent Gesture/Presentation was reliable but overall Parent Attention was not; this would have suggested that Gesture/Presentation plays a crucial role in predicting HSP scores; the present finding simply suggests that deictic gestures to a referent and manual presentations of the referent are two of several behaviors that reflect parent attention to the referent more generally, and it is this general attention that predicts HSP scores.

2.4. Summary and discussion of study 1

Two key findings emerge from this study. First, consistent with past work using HSP, we find that highly informative, referentially transparent, examples of caregivers uttering common concrete nouns are rare in the speech that children hear at home. Only 18% of all vignettes were classified as highly informative, i.e., as being correctly guessed by at least 50% of naïve adult viewers. Second, these highly informative occurrences had characteristic dynamic properties, as identified by our trained coders. Continued presence of the target referent throughout the vignette was not predictive of referential transparency – rather, a sudden appearance of the target referent 2–3 s prior to when the word was uttered, and continued presence immediately thereafter, was associated with higher referential transparency as measured by HSP responses. Additionally and separately, a sudden shift in child or parent attention just prior to word utterance and continuing thereafter was also associated with higher referential transparency. Parent gesture offered suggestive evidence of a similar contribution and timing but did not have a significant effect when entered into a model with the other measures of attention.

These timing patterns converge satisfyingly with those observed in the object-play studies of Smith and other investigators conducted in the lab and using different methods of data generation and analysis and with different participant pools (e.g., Pereira et al., 2014; Yu & Smith, 2013). Shifts in attention just prior to, and sustained after, labeling play an important role in referential transparency. Consistent with other laboratory studies (e.g., Baldwin, 1991, 1993; Tomasello & Farrar, 1986), our multiple regression findings suggest that there are several paths to referent identification: attention captured by an object's sudden appearance, follow-in labeling in which a parent labels a referent that a child is already attending to, and redirection in which first parent and then child both attend to a labeled object. Yet, the present data also suggest that all instances of this sort are rare in the home, implying perhaps that only a small percentage of word occurrences actually contribute to early vocabulary growth – a topic to which we return in the General Discussion. Before doing so, we offer Study 2, which solidifies our understanding of the timing characteristics of referentially transparent speech acts, offering, at least on a generous reading, a basis for cause-effect interpretation of these findings that goes beyond correlation.

3. Study 2: Temporal precision in reading referential intentions

Study 1 suggested that precise timing patterns are crucial to reading referential intent in natural parent–child interactions. Now we directly examine the learning value of these patterns by deliberately disrupting them. In three HSP experiments (2a & 2b & 2c) we surreptitiously moved the audible beep 1 s to 4 s away from its actual word occurrence within all the otherwise informative (HI) vignettes. This manipulation allows us to examine the language-cue timing relationship within the same set of videos, while controlling all other factors. Study 2a examined −1 s, 0 s, and +1 s offsets; Study 2b examined −2 s, 0 s, and +2 s offsets; Study 2c examined 4 s, 0 s, and +4 s offsets. Such timing disruption, in the limit, must have an effect. But the point of our brief timing disruptions is to gauge the notion of "temporal contiguity" more precisely than heretofore. Is the viewer sensitive to brief disruptions as he is, by analogy, to the sound track and mouth motion coming unglued as in badly engineered video presentations?

Relatedly, it is entirely possible that HI vignettes have behaviors that generally support referent identification, and that the temporal properties of these behaviors arise coincidentally in these videos. If so, manipulation of beep-offset should have little effect on HSP accuracy. But if the simultaneity of social-attentive behaviors relative to actual word utterance are crucial for inferring referential intent, we would expect drops in performance when the beeps are displaced from their original temporal positions.

3.1. Method

3.1.1. Participants

Ninety-four native English-speaking University of Pennsylvania undergraduates (36 in Exp. 2a, 29 in Exp. 2b, and 29 in Exp. 2c) participated for course credit.

3.1.2. Materials

We selected 27 HI and 10 LI vignettes as target and filler items, respectively (LI vignettes were included to keep the task challenging and provide a stimulus set more representative of parent–child interactions). All vignettes had only a single instance of the target word, at approximately 30 s. Targets were 20 nouns (Table 1), one from each of 27 families, and each target word occurred no more than twice. The patterns of social-attentive cues over time for this subset of HI vignettes are representative of the HI vignettes graphed in Fig. 2.

Three versions of each target were created per experiment, displacing the beep in target vignettes by −1 s, 0 s, and +1 s in Study 2a, −2 s, 0 s, and +2 s in Study 2b, and −4 s, 0 s, and +4 s in 2c. There were three experimental lists per experiment, each with the same fixed random order. In List 1, each target was randomly assigned to one of three beep-offset conditions; then conditions were rotated across the lists. Fillers were randomly interspersed. Reverse-order lists were also used.

3.1.3. Procedure

Participants were tested individually or in groups of 2–8. All participants watched all videos on either a projected screen or large screen plasma television in a conference room. A set of loudspeakers was used to project the sound of the beeps.

All participants read and signed a consent form before the start of the experiment. The experimenter then handed out a response sheet to each participant. On the response sheet were spaces to write down a guess for each of the 37 videos and a 5-point scale on which participants were to indicate their confidence in this guess. The experimenter also read aloud a set of detailed instructions stating that the videos they were about to view were

of parents and their children in the home; that the audio had been muted, but a single audible beep would identify the exact moment the parent had uttered a "mystery word," and that their job was to guess it.

Participants watched each video once. At the end of each video, they wrote down a single-word guess of what the parent in the video had uttered. This delayed-response (post-stimulus) procedure was adopted to ensure that participants had access to all information available in the entirety of each video at the point of making their guess. A 2-min break occurred after half of the videos had been shown. After completing all 37 videos, participants received both an oral and written debriefing.

3.1.4. Analyses

Following the procedure reported in Cartmill et al. (2013), all guesses that matched base morphemes of the target words were coded as correct, including diminutive and long/short alternates (i.e., telephone/phone, dog/doggie). Effect of beep offset on HSP accuracy was analyzed in three multilevel logit models (for Studies 2a, 2b, and 2c respectively), with crossed random intercepts and slopes included for Subjects and Items. The 0 s offset condition was the baseline.

3.2. Results

3.2.1. HSP accuracy

Effects of beep offset on HSP accuracy appear in Fig. 4. Accuracy of guesses dropped significantly when word onset was displaced by 4 s and even by 2 s. There was a slight asymmetry between negative and positive offsets, such that a −2 and −4 offsets was more detrimental to HSP accuracy than a +2 and +4 offsets. Offsets of −1 s/+1 s had no reliable effect on accuracy. Results of separate logit models for Studies 2a, 2b & 2c (Table 4) confirm these differences.

3.2.2. HSP confidence

HSP rating scores of confidence (from 1 to 5) were first standardized (*z*-scored) within each subject. We begin by noting an important fact: HSP observers are more confident about their correct guesses ($M = +0.31$) as compared to their incorrect guesses ($M = -0.44$), resulting in a highly significant effect of accuracy ($t(93) = 16.93$, $p < .01$) when collapsing across all three studies. This effect indicates that HSP observers have some implicit understanding of when their guess is correct. A similar effect of accuracy on confidence was observed by Medina et al. (2011, Supplemental Information), where, as in the present work and all other HSP studies that we know of, no explicit feedback was given to observers about the accuracy of their responses. Thus, for any given context, a correct guess makes greater sense to an observer than an incorrect one. Assuming this information is available to word learners, it can be used as implicit feedback regarding the accuracy of their referential hypothesis at that moment.

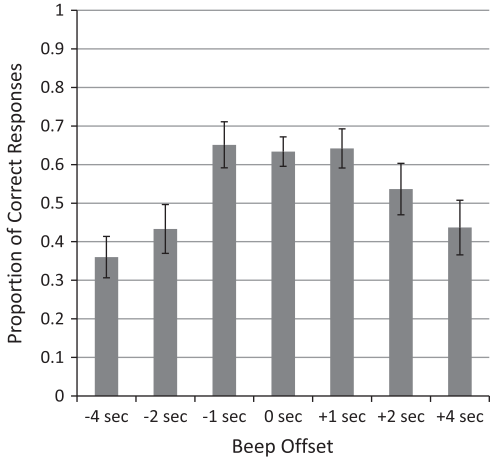

Fig. 4. Average proportion of correct responses (HSP accuracy) as a function of beep offset from actual word occurrence. Averages based on subject proportions. Error bars indicate 95% confidence intervals. 0 s offset data was collapsed over Studies 2a and 2b and 2c where the accuracy levels were 0.66, 0.59 and 0.64, respectively.

Table 4
Summary of fixed effects in mixed effects logit models predicting HSP accuracy with 0 s offset condition as baseline in Studies 2a, 2b, and 2c separately.

Study	Predictor	Coefficient	SE	Wald Z	*p*
2a	Intercept	0.79	0.23	3.51	<.001
	−1 s offset	−0.04	0.24	−0.16	–
	+1 s offset	0.03	0.31	0.09	–
2b	Intercept	0.52	0.29	1.75	–
	−2 s offset	−0.93	0.31	−2.97	<.01
	+2 s offset	−0.42	0.29	−1.422	–
2c	Intercept	0.74	0.25	2.94	<.01
	−4 s offset	−1.62	0.35	−4.63	<.001
	+4 s offset	−1.08	0.35	−3.06	<.01

Note: Results from three multi-level logit models predicting binary HSP Accuracy, with random intercept for Subjects and random intercept and offset condition as random slope for Items.

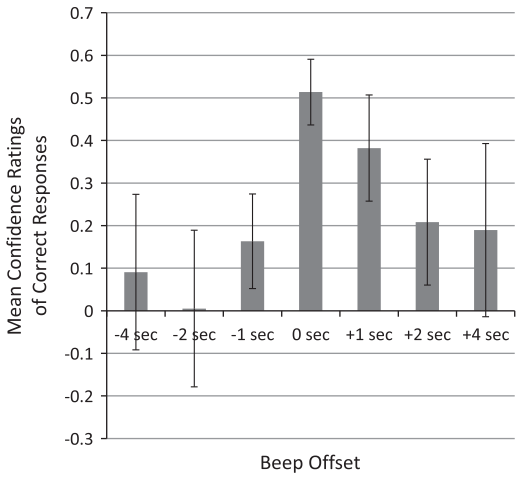

Fig. 5. Average normalized confidence ratings (*z*-scores within each subject) for correct HSP responses only as a function of beep offset from actual word occurrence. Averages based on subject means. Error bars indicate 95% confidence intervals. 0 s offset data was collapsed over Studies 2a and 2b and 2c where the average normalized confidence ratings were 0.434, 0.554 and 0.573, respectively.

A sub-analysis of confidence ratings from only correct guesses suggests that the timing between word onset and the visuo-social context is likely one source of feedback. In particular, Fig. 5 presents the effects of beep offset on HSP confidence for correct guesses only. Confidence dropped significantly when word onset was displaced even by 1 s. Note also there is a strong asymmetry between negative and positive offsets, such that a −1 offset is more detrimental to HSP confidence than a +1 offset. Results of separate linear models for Studies 2a, 2b & 2c (Table 5) confirm these differences. (Corresponding analyses of the ratings of incorrect guesses yielded no significant effects of beep offset, all *p*'s > .1.)

3.3. Discussion of study 2

Small perturbations in timing of the linguistic event (word onset) relative to context significantly reduce observers' ability

Table 5
Summary of fixed effects in multi-level models predicting normalized confidence ratings of correct responses with 0 s offset condition as baseline (Studies 2a, 2b, and 2c.).

Study	Predictor	Coefficient	SE	df	*t*-value	*p*
2a	Intercept	0.34	0.10	25.48	3.29	<.01
	−1 s offset	−0.22	0.09	62.50	−2.58	<.05
	+1 s offset	−0.00	0.08	38.87	0.02	–
2b	Intercept	0.44	0.12	26.63	3.56	<.01
	−2 s offset	−0.47	0.12	34.09	−3.83	<.001
	+2 s offset	−0.20	0.14	22.13	−1.46	–
2c	Intercept	0.51	0.13	27.31	3.99	<.001
	−4 s offset	−0.45	0.19	24.88	−2.41	<.05
	+4 s offset	−0.47	0.18	23.53	−2.64	<.05

Note: Three multi-level linear models predicting confidence ratings (correct trials only), with random intercept and slope for Subjects and for Items, with the exception of Study 2a, where converging model had no random slope for subject. The lmer-Test package to estimate *t*-values in multi-level model, estimating denominator degrees of freedom using "Satterthwaite" method.

and confidence in inferring referential intent. This importance of timing during reference (see also Gogate et al., 2000; Hollich et al., 2000; Jesse & Johnson, 2008) is reminiscent of the perception of causation linking two successive events. There, if the timing between the two events is increased beyond even 1/2 s, causality is less often ascribed (Leslie & Keeble, 1987; Michotte, 1946/63; Scholl & Nakayama, 2002). Relatedly, several prior investigations have shown that infants are sensitive to the timing of feedback in interpersonal interactions (e.g., Striano, Henning, & Stahl, 2006). The observed asymmetry in the effect of timing is consistent with the patterns in Study 1: the greatest changes in cues to referential intent occur just before, rather than after, word onset (Fig. 2); moving the beep early effectively causes these events to happen too late to be perceived as causally related to the linguistic event. As we discuss below, if word learners had access to this information it could be used to avoid incorrect referential hypotheses.

4. General discussion

4.1. Summary and contribution of present findings

We presented here an analysis of situational components of child–parent interactions to understand better what makes it possible for an observer (especially an infant) to reconstruct the speaker's intended referent for some simple words when uttered. The prior literature, as described in the Introduction, has richly demonstrated circumstances in which humans, adults and infants, can identify referents from the surrounding situational context. The task here was to find the loci of such referentially informative contexts as these might occur in parents' everyday conversations with their children. Overall, one can draw two conclusions from our findings. On the one hand, most (over 80%, in our sample) of everyday conversation, even restricting inquiry to the simplest (concrete nominal) words, is so ambiguous as to not support reconstruction of the speaker's intended referent (assuming that the HSP subjects are reasonable proxies in the relevant regards). But on the other hand, a small percentage of everyday usage appears to approximate the conditions found in laboratory studies, at least enough so that observers can guess the speaker's intended referent. The question, then, is how an observer, armed with human powers of inference, locates the islands of referential clarity in what must be, at least for a true novice, a sea of incomprehensible babble. This question motivated our analysis of parental speech to 14–18 month olds.

By coding social-attentive behaviors of parents and children in our videos and relating these codes to HSP accuracy (Study 1), we identified seven characteristics of specially informative scenarios that make them transparent to referential reconstruction: (1) increased likelihood that the target referent will appear, starting about 2 s before word onset; (2) increased Parent Attention to the target, sharply rising 1–3 s before word onset; (3) increased Parent Gesture/Presentation of the target one second before word onset; (4) increased Child Attention to the target beginning 3 s before word onset if not earlier; (5) decreased Parent and Child Attention to non-target referent objects starting at word onset and persisting about 8 s after word onset; (6) observations (2) through (5) still hold for the subset of vignettes in which the object is present throughout the entire video, indicating that observation (1) is not driving (2) through (5); and (7) co-incident attention of parent and child, though occurring at a low rate, occurs in patterns that are consistent with joint attention (generally) playing a central role in HI interactions.

These findings make three significant contributions. First, we have provided naturalistic validation of prior experimental work on the role of joint attention in reference identification (e.g., Baldwin, 1991, 1993; Jaswal, 2010; Nappa et al., 2009; Southgate et al., 2010; Tomasello & Farrar, 1986; Woodward, 2003) or observational studies of object play between parent and child (e.g., Harris et al., 1986; Pereira et al., 2014; Tomasello & Farrar, 1986; Tomasello & Todd, 1983; Tomasello et al., 1986). This past work suggested three ways that referent identification from extralinguistic information is accomplished: (a) capturing of attention by a salient object; (b) follow-in labeling in which the parent labels what is the current focus of the child's attention; and (c) redirected labeling in which the parent successfully redirects the child's attention to a labeled object via gesture or other evidence of a different attentional stance. We found evidence of similar patterns marking events of referential transparency in everyday usage in the home.

The second finding concerns the temporal dynamics of social-attentive behaviors in the wild, whose importance in referent identification has only recently been studied (e.g., Pereira et al., 2014). Study 2 provides especially useful documentation by showing directly that temporal disruption of word utterance with these social-attentive cues yields striking decrements in the ability to reconstruct the referent. Thus Study 2 more securely establishes a cause and effect relation between social-attentive cue timing and referent identification. Indeed the narrowness of the temporal window of this co-incident occurrence of word and event (a window of ±1 or 2 s, for accuracy and confidence level) is such as to suggest an analogy to the case of the attribution of physical causality when body 1 strikes body 2, and then body 2 moves, where for baby observers the interval must be circa 1/2 s (Leslie & Keeble, 1987; Michotte, 1946/63; Scholl & Nakayama, 2002). Study 1 and Study 2 identify some asymmetry in the timing of such behaviors, with attention persisting more after the word's occurrence, an observation also made in Pereira et al. (2014). This finding is worth further examination because it may suggest a role for sustained attention in memory encoding (see Pereira et al., 2014, for a discussion).

Third, the present findings reveal that information from the extralinguistic environment offers only rare moments of referential transparency in parent–child conversations at home, even for the concrete common nouns studied. Other lexical items (abstract nouns, most verbs, etc.) benefit even less from extralinguistic observation (e.g., Gillette et al., 1999; Snedeker & Gleitman, 2003), suggesting one important reason why early child vocabularies are over populated with the more concrete terms. All of this indicates that rare, highly informative events provide the initial

seed vocabulary for the learning of lexical items, a topic that we return to below in Section 4.3.

4.2. Logic and limitations of present work

Our findings are necessarily contingent on the power and validity of the assessment tools used. Therefore we now reconsider the logic and limitations of our measure of referential transparency: accuracy scores of HSP adults who guessed what the mother had said. In HSP, the experimental subjects are best thought of as "the crowd", a group of naïve individuals who, taken together, can in the presence of a mother/child dyad interacting provide a range of plausible guesses about what the mother was uttering. Though these participants are adults, it is assumed that they face essentially the same task that any learner confronts in the earliest moments of language learning (for a discussion, Gillette et al., 1999). Whether one has the linguistic sophistication of Noam Chomsky or the philosophical wisdom of Aristotle, still one has to glean on the first day in France that /ʃjɛn/ expresses the concept 'dog.' And if in Spain, /pɛrro/. The word–world connection has to be solved by every learner, as prerequisite to entry into further complexities of the language system. To this extent, any human who does not know a particular language, despite varying sophistication in every other regard, can serve as proxy for this question. Beyond this logical point, other evidence bolsters the validity of the HSP method. First, adults and children behave much alike in (suitably stripped down) versions of HSP presented to children (Medina et al., 2011; Piccin & Waxman, 2007). Moreover, consistent with the notion that HSP is measuring ease of referential identification for the children shown in the videos, Cartmill et al. (2013) found that HSP accuracy scores of these videos predicted the children's vocabulary scores three years later: Families whose parent–child interactions at 14- to 18-months were highly referentially transparent (as measured by HSP accuracy) have children whose vocabulary size is larger 3 years later at school entry than families whose interactions at 14- to 18-months were not as referentially transparent.[7]

Nevertheless the immediate experience (and cognition) of the adult HSP subjects is not the same as that of the infants in the videos. The differences here restrict and constrain how the present results should be interpreted. Most notably, the visual perspective and the task of the HSP participants differ from those of the child in the video. With respect to the visual perspective, there is indeed a growing literature identifying differences between the information available from a 1st person vs. 3rd person view of a task or interaction, some of which were discussed in our introductory remarks (Yoshida & Smith, 2008; see also Kretch, Franchak, & Adolph, 2014; Smith, Yu, Yoshida, & Fausey, 2015). Striking as these eye-view differences are, the question is how they impact the ability to read referential intent from these two camera angles. To the extent that this has been studied, comparisons of HSP responses from 1st- and 3rd-person cameras suggest only modest reductions in accuracy for 3rd-person cameras.

In particular, Yurovsky et al. (2013) compared HSP responses from the same parent child-interactions as recorded from a 1st-person head-mounted child camera vs. a fixed 3rd-person camera. As mentioned earlier, their genre was object play, and only parent utterances that labeled co-present objects were used as stimuli. As such, average HSP accuracy scores were overall considerably higher (58% correct) than most previous studies of random samples of concrete nouns (e.g., Cartmill et al., 2013; Gillette et al., 1999; Medina et al., 2011; and Study 1 results above). Most relevantly, however, Yurovsky et al. (2013) report that differences in HSP responses between camera angles are relatively small: average HSP accuracy was 58% for both 1st and 3rd person camera (identical and not statistically different, p. 960), and the 3rd person camera actually offered more "unambiguous" interactions in which all HSP participants (100%) correctly guessed the mother's word utterance (1st person: 10% unambiguous vignettes; 3rd person: 22% unambiguous vignettes; Fig. 1, p. 961) with very similar accuracy distributions across vignettes. The only HSP advantage for 1st over 3rd person cameras came when vignettes of low informativity were strung together in a study of cross-situational word learning: a 1st person camera offered greater improvements over a 3rd person camera (Fig. 3, p. 963). Thus, although one might expect large differences between a 1st and 3rd person camera in terms of an observer guessing what the mother had said, differences in this case appear to be minimal. The gaze following literature may offer an explanation: adult observers are good at judging what another person is attending to under live-action conditions in which head-turn and gaze-information are consistent and apparent (e.g., Doherty, Anderson, & Howeieson, 2009, Experiment 2, >90% accuracy). Adult HSP observers, to the extent that they use similar cues, should be able to assess attentional states of child and parent from a 3rd person view with some accuracy, and apparently do so at rates similar to a 1st-person child view.[8]

The gaze-following literature does, however, suggest that adult HSP observers likely accomplish referent identification in different ways than the children in the videos because the adult HSP task is the result of an explicit meta-cognitive judgment task whereas the child's task is an implicit one. Indeed, 2- and 3-year old children are worse than adults at making explicit judgments of what a person is looking at, but nevertheless do accurately follow the gaze of another individual (e.g., Doherty & Anderson, 1999; Doherty et al., 2009). And infants and toddlers make other implicit decisions that support the reading of intent (e.g., Kovács, Téglás, & Endress, 2010; Onishi & Baillargeon, 2005; Surian, Caldi, & Sperber, 2007). Thus the end result is often the same for child and adult: attentional focus on the speaker's intended referent is accomplished by the child without necessarily being able to explicitly describe that state to another in a judgment task (Doherty et al., 2009). It is likely then that although HSP codes might slightly underestimate the incidence of referential transparency, our codes offer a reasonable assessment of this property of the parent–child interchange.

There are several other obvious differences between HSP subjects and infants. Adults and infants differ in their knowledge of the world, of typological differences between possible words in different languages (e.g., Bowerman & Levinson, 2001; Naigles & Terrazas, 1998; Slobin, 2008), and biases that stem from culture or adult cognition. For instance, there is evidence that younger children exhibit biases about what constitutes a word meaning that change with age (e.g., Landau, Smith, & Jones, 1988, 1992; Markman, 1990). These differences could artificially increase or decrease the accuracy of HSP subjects, suggesting some caution in interpreting our results. Yet, there is also ample evidence of conceptual sophistication in children of the age of interest (e.g., Csibra, Bíró, Koós, & Gergely, 2003; Gergely & Csibra, 2003; Scott &

[7] The range of HSP transparency varies across families from a low of only 4 percent to a high of 38 percent in this sample, revealing a striking disparity among families as to how much adults are "speaking to" their children rather than "speaking at" them. Notably this measure of the quality of input (as opposed to sheer quantity) is not significantly related to SES but appears to be an individual family characteristic.

[8] An additional worry about the generality of HSP findings comes from the reasonable complaint that these are usually judgments by undergraduates at an elite American university. And of course, the children themselves, as we have described, come from varied SES backgrounds. However, the HSP experiment that we used in Study 1 (as collected by Cartmill et al., 2013) included participants from a local city college in which a significant proportion are first in their family to attend college. No differences in HSP accuracy between these different college populations were found, nor did the reported maternal education of HSP participants predict HSP guessing accuracy (Cartmill et al., 2013, *Supporting Information*).

Baillargeon, 2013; Setoh, Wu, Baillargeon, & Gelman, 2013), suggesting conceptual overlap between HSP participants and the children in the videos. We believe, however, that where differences do emerge they will not greatly affect our general finding, namely that reading referential intent from extralinguistic information alone is a difficult task for everyone, successful only on rare special occasions. This finding points to a class of learning mechanisms that heavily filter the input, with most (referentially opaque) incidences of word occurrence not even entering into the search for meaning.

4.3. Implications for word learning

The present findings comport well with theorizing and observations about vocabulary acquisition from our laboratories (Koehne, Trueswell, & Gleitman, 2013; Medina et al., 2011; Trueswell, Medina, Hafri, & Gleitman, 2013). There we find that adult learners (and child 2-to-3-yr-old learners, see Woodard, Gleitman, & Trueswell, in press) have a strong bias to identify just one plausible referent when hearing a novel word, with only a single meaning being carried forward from any given learning instance. The result is that low-informative contexts, i.e., contexts that offer many candidate referents, do not flood the learner with many (typically incorrect) hypotheses, as would be the case for a fully "cross-situational" word learner who extracted and retained all plausible candidate referents from each learning instance. A more focused, single-hypothesis learner would retain only a single guess from a low-informative instance, which although likely incorrect, would not flood the learner with additional incorrect hypotheses.

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, Study 2 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. We identified one plausible candidate for such feedback: the timing of visuo-social cues to reference relative to word onset. In particular, 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 (Fig. 5) and a different guess may be posited (Fig. 4).[9]

We propose 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 by the learner. 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.

Although we argue that epiphany moments do occur in early word learning, these moments were necessarily preceded by several important linguistic developments in other domains. Via sheer exposure to speech, infants first develop candidate syllable and word forms (Saffran, Aslin, & Newport, 1996, and e.g., Bortfeld, Morgan, Golinkoff, & Rathbun, 2005; Hay, Pelucchi, Estes, & Saffran, 2011), and know relevant syntactic category information (Mintz, Newport, & Bever, 2002) and prosodic information (Johnson & Jusczyk, 2001; Thiessen & Saffran, 2003), all of which contribute to the word learning process even at early stages (e.g., Hollich et al., 2000; Waxman & Booth, 2001), with some word meanings perhaps being learned as early as 6 months of age (Bergelson & Swingley, 2012). Our contention here is that there are moments when such knowledge aligns with the referent world in precise ways to make referent identification and the learning of word meanings possible, and that such informative events are rare (see also, e.g. Nelson, Welsh, Camarata, Heimann, & Tjus, 2001).

Relatedly, as embedded in the very choice of materials for the analyses presented here, learning "from context," even from precisely timed and focused context, works almost solely for a very limited subset of common words, namely, the basic-level "imageable" nominal vocabulary that dominates knowledge in the first and second years of life. But this first primitive vocabulary, limited though it is, plays two crucial roles in the life of the child entering the human interpersonal environment. First, with a little good will and expansive communicative intent, meaningful conversation between infants and caretakers becomes increasingly successful during this time period, supporting social binding and increased responsiveness to social cues to speaker intent (e.g., Golinkoff, 1986; Golinkoff & Hirsch-Pasek, 2006; Kinzler, Dupoux, & Spelke, 2007). Second, these initial words enable the learner to discover the canonical forms of sentences in the exposure language, e.g., that English is SVO (cf., Bever, 1970). This primitive phrase-structural and morphological knowledge, in turn, underpins a new phase of learning in which the structural and distributional properties of speech in linguistic context enables the learning of words from every linguistic and conceptual category, in all their abstraction and variety (Gleitman et al., 2005). Vocabulary growth begins its rapid climb and achieves its apogee of rate and category breadth only when this second, structure sensitive, stage is reached. While reference-finding remains central to vocabulary growth over the lifespan, the scaffolding provided by early lexical acquisition crucially transmutes the learning procedure itself from word-to-world pairing, as we have investigated it here, to structure-to-world pairing, or "syntactic bootstrapping."

[9] It would be possible for a fully-cross-situational word learner to take advantage of similar feedback, by evaluating how a range a possible referential hypotheses from a given instance fit expected timing characteristics – or perhaps more simply, by down-weighting the effects of a referential hypothesis when it appears in a context that also offers many other hypotheses (as would be the case in low-informative contexts). But here we note an important observation about confidence from Medina et al. (2011): people were just as confident about a correct guess from a low informative context as a correct guess from a high informative context, even though a low informative context would likely offer many more competing hypotheses. This finding suggests that learners are not evaluating the "spread" of competing referential hypotheses and instead are evaluating the quality of just one hypothesis.

Acknowledgments

Research supported by Supplemental American Recovery and Reinvestment Act (ARRA) Grant (3R01HD037507-11S1) awarded to J.C.T. and L.R.G., as part of the National Institute of Child Health and Human Development Grant 3-R01HD37507 (P.I.s: L.G. and J.T.). Support also from NICHD Grant P01HD40605 to S.G-M. Research was approved by institutional review boards at UPenn, La Salle and Chicago.

Appendix A

See Table A.1.

Table A.1
Examples of HSP Responses from Study 1. (Number of responses of each type in parentheses.)

1. Vignette type: HI; word uttered: "dog"; HSP responses: dog (9), duck (2), cat (1), horse (1)
2. Vignette type HI; Word uttered: "hair", HSP responses: hair (8), handsome (1), head (1), mess (1), silly (1), sit (1)
3. Vignette type: MI; Word uttered "book"; HSP responses: book (7), done (2), read (2), where (2), going (1), toy (1), yellow (1)
4. Vignette type: MI; word uttered "nose", HSP responses: nose (5), glasses (4), eye (3), face (2), gentle (1), touch (1)
5. Vignette type: LI; word uttered "dog", HSP responses: couch (3), here (3), go (2), stand (2), again (1), chair (1), come (1), pacifier (1)
6. Vignette type: LI; word uttered "bird", HSP responses: go (3), room (3), carry (1), door (1), here (1), inside (1), kitchen (1), let's (1), outside (1), yes (1)

Appendix B

For the 351 Study 1 vignettes in which the child did not know the target word, Generalized Logit Models were used to test the correlation between presence/absence of social-attentive cues and HSP accuracy. Significant correlations are reported in Table B.1.

Table B.1
Reliable chi-square correlations between HSP accuracy (E-logit) and presence/absence of social-attentive cues, after adjusting for multiple tests using Bonferroni's method (6 tests per 31 s, resulting in 186 tests in total, thus only p's < 0.00027 are reported).

Cue	Time (s) relative to word onset	Chi-square correlation	Beta coefficient
Referent Presence	−2	13.66	0.26
	−1	20.18	0.34
	0	35.06	0.50
	+1	60.03	0.72
	+2	45.51	0.57
	+3	35.03	0.47
	+4	22.66	0.35
	+5	24.01	0.37
	+6	20.39	0.33
	+7	18.29	0.31
	+8	16.73	0.29
	+9	21.17	0.38
Parent Gest./Presentation of Referent	−1	15.84	0.49
	0	17.21	0.44
	+1	23.22	0.54
	+2	25.16	0.61
	+3	14.53	0.52
	+4	16.82	0.53
Parent Attention to Referent	−3	16.10	0.42
	−2	26.20	0.47
	−1	44.89	0.55
	0	49.91	0.54
	+1	54.61	0.57
	+2	53.74	0.57
	+3	40.68	0.51
	+4	43.45	0.53
	+5	31.69	0.46
	+6	32.50	0.47
	+7	29.92	0.47
	+8	26.98	0.46
	+9	21.03	0.47
Child Attention to Referent	−3	21.61	0.38
	−2	30.27	0.44
	−1	56.21	0.60
	0	60.57	0.62
	+1	62.76	0.62
	+2	54.48	0.56
	+3	56.03	0.59
	+4	63.71	0.64
	+5	48.91	0.58
	+6	47.79	0.59
	+7	37.37	0.53
	+8	28.75	0.44
	+9	15.71	0.36
Parent Attention to Non-Referents	+1	−14.90	−0.28
Child Attention to Non-Referents	0	−17.86	−0.29
	+1	−17.51	−0.28
	+2	−20.75	−0.31
	+3	−34.28	−0.41
	+4	−24.75	−0.34
	+5	−21.65	−0.32
	+6	−22.18	−0.32
	+7	−24.65	−0.34
	+8	−20.80	−0.31

Appendix C

See Fig. C.1.

Appendix D

For the 351 Study 1 vignettes in which the child did not know the target word, multiple regressions were conducted on a second-by-second basis (−20 s +10 s from word onset). Specifically, 31 General Linear Models (1 per second) were constructed such that E-logit HSP Accuracy scores were predicted by: Referent Presence (no/yes), Parent Gesture/Presentation of Referent (no/yes), Parent Attention to Referent (no/yes), Child Attention to Referent (no/yes) and the interaction between Child and Parent Attention to Referent. Model results presented below are from the models that had significantly better fits than empty models with no fixed effects, based on a chi-square test of the

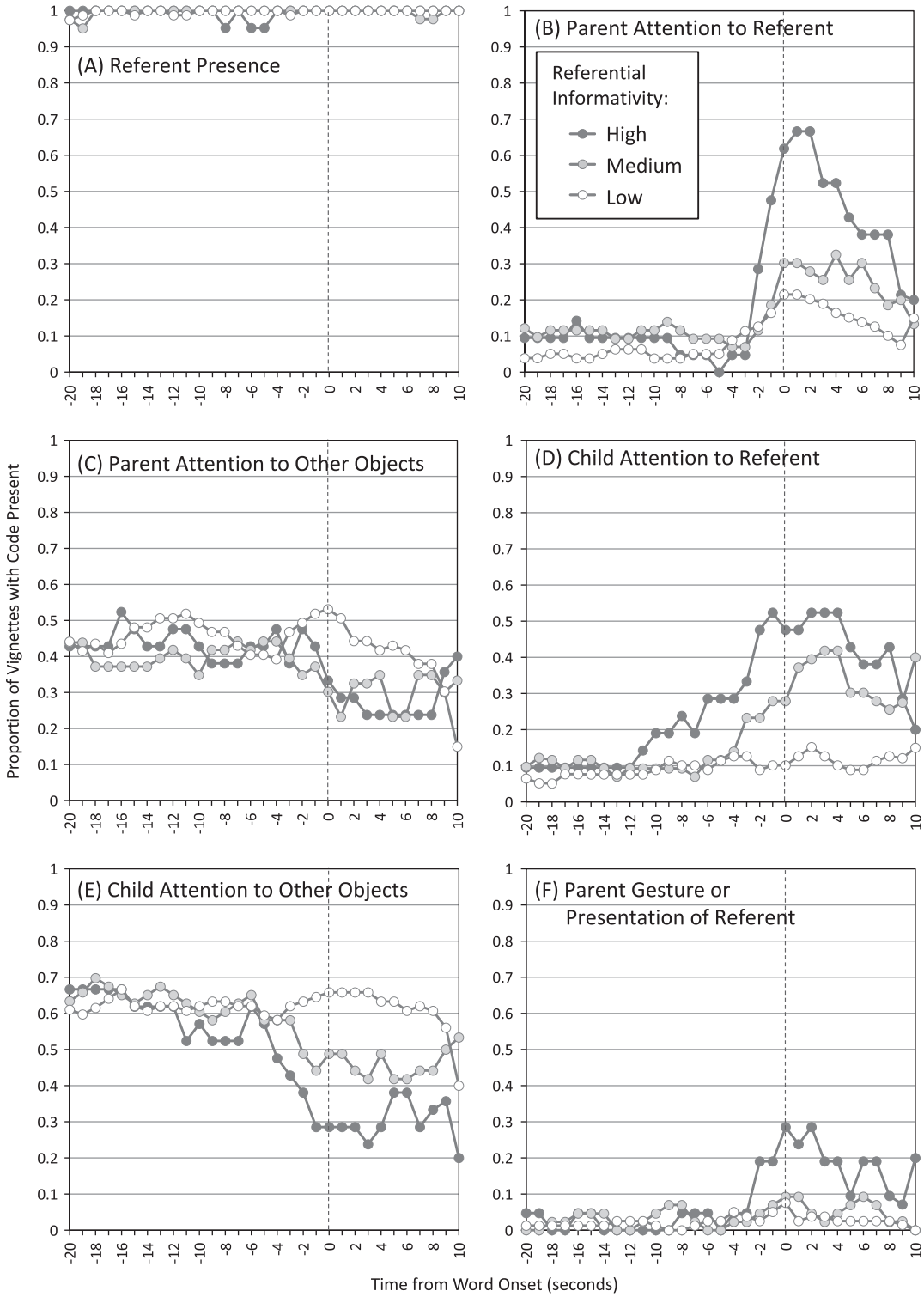

Fig. C.1. Proportion of vignettes with social-attentive cues coded as present (*y*-axis) in relation to word onset (dotted line on *x*-axis) in a subset of Study 1 vignettes where the referent is present 95% or more of the time ($n = 143$). Vignettes are grouped according to HSP informativity (High, Middle & Low).

change in −2 restricted log likelihood (Steiger, Shapiro, & Browne, 1985). This test was deemed significant only after correcting for the number tests using Bonferroni's method (31 models, resulting in a significance level of $p < .0016$). Co-efficient *t*-tests below are tests against the normal distribution (with $p < 0.01$ marked in **bold**). It is important to note that with the exception of the model of the 10th second, the models did not violate assumptions of collinearity in predictors, in that the square root of the variable inflation factor for each predictor did not exceed 2. This means that these necessarily correlated predictors nevertheless can be added to the same multiple regression (see Table D.1).

Table D.1
Reliable multiple regression linear models on second by-second basis from Study 1.

Second	Measure	Fixed effects					
		Intercept	Ref. Pres.	Par. Gest. Presentation	Par. Att. to Ref.	Ch. Att. to Ref.	Par. X Ch. Att. to Ref.
−5 s	Coeff.	−2.39	0.42	−1.48	**1.66**	0.60	−1.04
	S.E.	0.13	0.18	1.25	**0.51**	0.28	0.80
	t-value	−17.78	2.33	−1.18	**3.24**	2.14	−1.30
	p-value	0.0000	0.0203	0.2389	**0.0013**	0.0327	0.1947
−4 s	Coeff.	−2.37	0.33	−0.96	**1.43**	**0.76**	−0.75
	S.E.	0.14	0.18	0.69	**0.50**	**0.27**	0.70
	t-value	−17.37	1.82	−1.39	**2.84**	**2.77**	−1.08
	p-value	0.0000	0.0704	0.1642	**0.0048**	**0.0059**	0.2811
−3 s	Coeff.	−2.44	0.31	0.07	**1.52**	**1.01**	−1.30
	S.E.	0.13	0.18	0.64	**0.45**	**0.26**	0.62
	t-value	−18.12	1.76	0.11	**3.39**	**3.94**	−2.11
	p-value	0.0000	0.0794	0.9116	**0.0008**	**0.0001**	0.0356
−2 s	Coeff.	−2.53	0.32	0.18	**1.46**	**1.16**	−1.19
	S.E.	0.13	0.18	0.56	**0.34**	**0.26**	0.55
	t-value	−18.84	1.85	0.32	**4.28**	**4.45**	−2.15
	p-value	0.0000	0.0655	0.7517	**<0.0001**	**<0.0001**	0.0324
−1 s	Coeff.	−2.69	0.26	0.23	**1.58**	**1.63**	**−1.45**
	S.E.	0.13	0.17	0.38	**0.30**	**0.24**	**0.43**
	t-value	−20.47	1.55	0.59	**5.36**	**6.75**	**−3.35**
	p-value	0.0000	0.1220	0.5556	**<0.0001**	**<0.0001**	**0.0009**
0 s	Coeff.	−2.82	0.39	0.16	**1.30**	**1.65**	**−1.14**
	S.E.	0.14	0.18	0.32	**0.27**	**0.26**	**0.40**
	t-value	−20.87	2.23	0.51	**4.85**	**6.45**	**−2.85**
	p-value	0.0000	0.0261	0.6121	**<0.0001**	**<0.0001**	**0.0047**
+1 s	Coeff.	−3.04	**0.70**	0.60	**1.16**	**1.51**	**−1.08**
	S.E.	0.13	**0.17**	0.32	**0.25**	**0.23**	**0.38**
	t-value	−23.07	**4.06**	1.90	**4.57**	**6.51**	**−2.88**
	p-value	0.0000	**0.0001**	0.0588	**<0.0001**	**<0.0001**	**0.0043**
+2 s	Coeff.	−2.94	**0.58**	0.77	**1.24**	**1.36**	**−1.02**
	S.E.	0.13	**0.17**	0.34	**0.26**	**0.23**	**0.39**
	t-value	−22.51	**3.42**	2.24	**4.82**	**6.00**	**−2.64**
	p-value	0.0000	**0.0007**	0.0260	**<0.0001**	**<0.0001**	**0.0087**
+3 s	Coeff.	−2.81	**0.46**	0.53	**1.16**	**1.49**	−0.91
	S.E.	0.13	**0.17**	0.40	**0.26**	**0.23**	0.41
	t-value	−21.74	**2.67**	1.35	**4.49**	**6.44**	−2.22
	p-value	0.0000	**0.0079**	0.1788	**<0.0001**	**<0.0001**	0.0274
+4 s	Coeff.	−2.73	0.24	0.55	**1.27**	**1.70**	−0.97
	S.E.	0.13	0.16	0.37	**0.26**	**0.23**	0.41
	t-value	−21.77	1.48	1.48	**4.92**	**7.52**	−2.38
	p-value	0.0000	0.1390	0.1398	**<0.0001**	**<0.0001**	0.0177
+5 s	Coeff.	−2.65	0.39	0.15	**0.94**	**1.41**	−0.44
	S.E.	0.13	0.17	0.41	**0.29**	**0.26**	0.46
	t-value	−20.30	2.25	0.37	**3.22**	**5.43**	−0.96
	p-value	0.0000	0.0249	0.7122	**0.0014**	**0.0000**	0.3363
+6 s	Coeff.	−2.61	0.33	0.29	**1.04**	**1.65**	−0.97
	S.E.	0.13	0.17	0.39	**0.30**	**0.28**	0.47
	t-value	−20.16	1.94	0.74	**3.49**	**5.86**	−2.07
	p-value	0.0000	0.0532	0.4576	**0.0005**	**0.0000**	0.0392
+7 s	Coeff.	−2.54	0.32	0.13	**1.24**	**1.48**	−1.12
	S.E.	0.13	0.17	0.44	**0.34**	**0.29**	0.51
	t-value	−19.90	1.85	0.30	**3.61**	**5.07**	−2.21
	p-value	0.0000	0.0657	0.7622	**0.0004**	**0.0000**	0.0280
+8 s	Coeff.	−2.52	0.33	−0.29	**1.51**	**1.24**	−1.31
	S.E.	0.13	0.18	0.50	**0.35**	**0.29**	0.52
	t-value	−19.17	1.84	−0.58	**4.36**	**4.31**	−2.50
	p-value	0.0000	0.0663	0.5636	**<0.0001**	**<0.0001**	0.0130
+9 s	Coeff.	−2.61	**0.53**	0.59	**1.36**	**0.86**	−1.30

(*continued on next page*)

Table D.1 (*continued*)

Second	Measure	Fixed effects					
		Intercept	Ref. Pres.	Par. Gest. Presentation	Par. Att. to Ref.	Ch. Att. to Ref.	Par. X Ch. Att. to Ref.
	S.E.	0.14	**0.19**	0.56	**0.40**	**0.30**	0.58
	t-value	−18.11	**2.77**	1.05	**3.44**	**2.89**	−2.24
	p-value	0.0000	**0.0059**	0.2943	**0.0007**	**0.0042**	0.0261
+10 s	Coeff.	−2.45	0.56	2.64	**−0.96**	1.03	0.71
	S.E.	0.26	0.35	1.02	**0.93**	0.60	1.14
	t-value	−9.59	1.60	2.59	**−1.03**	1.71	0.63
	p-value	0.0000	0.1137	0.0112	**0.3041**	0.0907	0.5329

References

Akhtar, N., & Gernsbacher, M. A. (2007). Joint attention and vocabulary development: A critical look. *Language and Linguistic Compass, 1*(3), 195–207.

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

Baldwin, D. A. (1993). Infants' ability to consult the speaker for clues to word reference. *Journal of Child Language, 20*, 395–418.

Baldwin, D. A. (1995). Understanding the link between joint attention and language. In C. Moore & P. J. Dunham (Eds.), *Joint attention: Its origins and role in development* (pp. 131–158). Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc..

Baldwin, D. A., & Tomasello, M. (1998). Word learning: A window on early pragmatic understanding. In E. V. Clark (Ed.), *The proceedings of the twenty-ninth annual child language research forum* (pp. 3–23). Chicago, IL: Center for the Study of Language and Information.

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, 109*, 3253–3258.

Bever, T. G. (1970). The cognitive basis for linguistic structures. In J. R. Hayes (Ed.), *Cognition and the development of language* (pp. 280–312). New York: Wiley.

Bloom, P. (2002). Mindreading, communication, and the learning of the names for things. *Mind and Language, 17*, 37–54.

Bortfeld, H., Morgan, J., Golinkoff, R., & Rathbun, K. (2005). Mommy and me: Familiar names help launch babies into speech stream segmentation. *Psychological Science, 16*, 298–304.

Bowerman, M., & Levinson, S. C. (2001). *Language acquisition and conceptual development* (Vol. 3). Edited volume. Cambridge, UK: Cambridge University Press.

Brown, R. (1973). *A first language: The early years*. Cambridge, MA: Harvard University Press.

Brown-Schmidt, S., & Tanenhaus, M. K. (2008). Real-time investigation of referential domains in unscripted conversation: A targeted language game approach. *Cognitive Science, 32*(4), 643–684.

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

Cartmill, E. A., Armstrong, B. F., Gleitman, L. R., Goldin-Meadow, S., Medina, T. N., & Trueswell, J. C. (2013). Quality of early parent input predicts child vocabulary 3 years later. *Proceedings of the National Academy of Sciences, 110*(28), 11278–11283.

Chomsky, C. (1969). *The acquisition of syntax in children from 5–10*. Cambridge, MA: MIT Press.

Csibra, G., Bíró, S., Koós, O., & Gergely, G. (2003). One-year-old infants use teleological representations of actions productively. *Cognitive Science, 27*(1), 111–133.

Doherty, M. J., & Anderson, J. R. (1999). A new look at gaze: Preschool children's understanding of eye direction. *Cognitive Development, 14*, 549–571.

Doherty, M. J., Anderson, J. R., & Howeieson, L. (2009). The rapid development of explicit gaze judgment ability at 3 years. *Journal of Experimental Child Psychology, 104*, 296–312.

Gergely, G., & Csibra, G. (2003). Teleological reasoning in infancy: The naıve theory of rational action. *Trends in Cognitive Sciences, 7*(7), 287–292.

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

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

Gogate, L. J., Bahrick, L. E., & Watson, J. D. (2000). A study of multimodal motherese: The role of temporal synchrony between verbal labels and gestures. *Child Development, 71*(4), 878–894.

Goldin-Meadow, S., Levine, S. C., Hedges, L. V., 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.

Golinkoff, R. M. (1986). "I beg your pardon?": The preverbal negotiation of failed messages. *Journal of Child Language, 13*(3), 455–477.

Golinkoff, R. M., & Hirsch-Pasek, K. (2006). Baby wordsmiths: From associationist to social sophistication, 2006. *Current Directions in Psychological Science, 15*, 30–33.

Grice, H. P. (1989). *Studies in the way of words*. Cambridge, MA: Harvard University Press.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics 3: Speech acts*. New York: Seminar Press.

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.

Hay, J. F., Pelucchi, B., Estes, K. G., & Saffran, J. R. (2011). Linking sounds to meanings: Infant statistical learning in a natural language. *Cognitive Psychology, 63*(2), 93–106.

Hollich, G. J., Hirsh-Pasek, K., Golinkoff, R. M., Brand, R. J., Brown, E., Chung, H. L., ... Bloom, L. (2000). Breaking the language barrier: An emergentist coalition model for the origins of word learning. *Monographs of the Society for Research in Child Development*, 1–135.

Hume, D. (1902). An enquiry concerning human understanding. In *An enquiry concerning human understanding and concerning the principles of morals*. Oxford: Clarendon Press (Original work published in 1748).

Jaswal, V. K. (2010). Believing what you're told: Young children's trust in unexpected testimony about the physical world. *Cognitive Psychology, 61*, 248–272.

Jesse, A., & Johnson, E. K. (2008). Audiovisual alignment in child-directed speech facilitates word learning. In *Proceedings of the international conference on auditory-visual speech processing* (pp. 101–106). Adelaide, Australia: Causal Productions.

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

Kinzler, K. D., Dupoux, E., & Spelke, E. S. (2007). The native language of social cognition. *The Proceedings of the National Academy of Sciences, 104*, 12577–12580.

Koehne, J., Trueswell, J. C., & Gleitman, L. R. (2013). Multiple proposal memory in observational word learning. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th annual meeting of the Cognitive Science Society* (pp. 805–810). Austin, TX: Cognitive Science Society.

Kovács, Á. M., Téglás, E., & Endress, A. D. (2010). The social sense: Susceptibility to others' beliefs in human infants and adults. *Science, 330*(6012), 1830–1834.

Kretch, K. S., Franchak, J. M., & Adolph, K. E. (2014). Crawling and walking infants see the world differently. *Child Development, 85*(4), 1503–1518.

Landau, B., & Gleitman, L. (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.

Landau, B., Smith, L. B., & Jones, S. S. (1992). Syntactic context and the shape bias in children's and adults' lexical learning. *Journal of Memory & Language, 31*(6), 807–825.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174.

Lausberg, H., & Sloetjes, H. (2009). Coding gestural behavior with the NEUROGES-ELAN system. *Behavior Research Methods, Instruments, & Computers, 41*(3), 841–849.

Lawson, K. R., & Ruff, H. A. (2004). Early focused attention predicts outcome for children born prematurely. *Journal of Developmental & Behavioral Pediatrics, 25*(6), 399–406.

Leslie, A. M., & Keeble, S. (1987). Do six-month old infants perceive causality? *Cognition, 25*, 265–288.

Lidz, J., Waxman, S., & Freedman, J. (2003). What infants know about syntax but couldn't have learned: Experimental evidence for syntactic structure at 18 months. *Cognition, 89*(3), 295–303.

Lyons, C. (1999). *Definiteness*. Cambridge, UK: Cambridge University Press.

Markman, E. M. (1990). Constraints children place on word meanings. *Cognitive Science, 14*, 154–173.

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

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

Mintz, T. H., Newport, E. L., & Bever, T. G. (2002). The distributional structure of grammatical categories in speech to young children. *Cognitive Science, 26*(4), 393–424.

Moore, C., & Dunham, P. (1995). *Joint attention: Its origins and role in development*. New York: Psychology Press.

Naigles, L. (1990). Children use syntax to learn verb meanings. *Journal of Child Language, 17*(2), 357–374.
Naigles, L. R., & Terrazas, P. (1998). Motion verb generalizations in English and Spanish: Influences of language and syntax. *Psychological Science, 9*, 363–369.
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.
Nelson, K. E., Welsh, J., Camarata, S., Heimann, M., & Tjus, T. (2001). A rare event transactional dynamic model of tricky mix conditions contributing to language acquisition and varied communicative delays. In K. E. Nelson, A. Koc, & C. Johnson (Eds.). *Children's language* (Vol. 11). Hillsdale, NJ: Erlbaum.
Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science, 308*(5719), 255–258.
Pereira, A. F., Smith, L. B., & Yu, C. (2014). A bottom-up view of toddler word learning. *Psychonomic Bulletin & Review, 21*(1), 178–185.
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.
Pruden, S. M., Hirsh-Pasek, K., Golinkoff, R., & Hennon, E. A. (2006). The birth of words: Ten-month-olds learn words through perceptual salience. *Child Development, 77*, 266–280.
Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science, 274*(5294), 1926–1928.
Scholl, B. J., & Nakayama, K. (2002). Causal capture: Contextual effects on the perception of collision events. *Psychological Science, 13*(6), 493–498.
Scott, R. M., & Baillargeon, R. (2013). Do infants really expect agents to act efficiently? A critical test of the rationality principle. *Psychological Science, 24*(4), 466–474.
Setoh, P., Wu, D., Baillargeon, R., & Gelman, R. (2013). Young infants have biological expectations about animals. *Proceedings of the National Academy of Sciences, 110* (40), 15937–15942.
Slobin, D. I. (2008). The child learns to think for speaking: Puzzles of crosslinguistic diversity in form-meaning mappings. In T. Ogura & H. Kobayashi, et al. (Eds.). *Studies in language sciences* (Vol. 7, pp. 3–22). Tokyo: Kurosio Publishers.
Smith, L. B., Colunga, E., & Yoshida, H. (2010). Knowledge as process: Contextually cued attention and early word learning. *Cognitive Science, 34*(7), 1287–1314.
Smith, L. B., Yu, C., Yoshida, H., & Fausey, C. M. (2015). Contributions of head-mounted cameras to studying the visual environments of infants and young children. *Journal of Cognition and Development, 16*(3), 407–419.
Smith, L. B., Yu, C., & Pereira, A. F. (2011). Not your mother's view: The dynamics of toddler visual experience. *Developmental Science, 14*(1), 9–17.
Snedeker, J., & Gleitman, L. R. (2003). Why it is hard to label our concepts. In S. Waxman & G. Hall (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.
Steiger, J. H., Shapiro, A., & Browne, M. W. (1985). On the multivariate asymptotic distribution of sequential chi-square statistics. *Psychometrika, 50*, 253–264.
Striano, T., Henning, A., & Stahl, D. (2006). Sensitivity to interpersonal timing at 3 and 6 months of age. *Interaction Studies, 7*(2), 251–271.
Surian, L., Caldi, S., & Sperber, D. (2007). Attribution of beliefs by 13-month-old infants. *Psychological Science, 18*(7), 580–586.
Thiessen, E. D., & Saffran, J. R. (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.
Tomasello, M., & Farrar, M. J. (1986). Joint attention and language. *Child Development, 57*(6), 1451–1463.
Tomasello, M., Mannle, S., & Kruger, A. C. (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.
Tomasello, M. (1995). Pragmatic contexts for early verb learning. In M. Tomasello & W. Merriman (Eds.), *Beyond names for things: Young children's acquisition of verbs*. Lawrence Erlbaum.
Trueswell, J. C., Medina, T. N., Hafri, A., & Gleitman, L. R. (2013). Propose but verify: Fast mapping meets cross-situational word learning. *Cognitive Psychology, 66*(1), 126–156.
Vlach, H. A., & Sandhofer, C. M. (2014). Retrieval dynamics and retention in cross-situational statistical word learning. *Cognitive Science, 38*(4), 757–774.
Waxman, S. R., & Booth, A. E. (2001). Seeing pink elephants: Fourteen-month-olds' interpretations of novel nouns and adjectives. *Cognitive Psychology, 43*(3), 217–242.
Woodard, K., Gleitman, L. R., & Trueswell, J. C. (in press). Two-and three-year-olds track single meaning during word learning: Evidence for propose-but-verify. *Language Learning and Development*.
Woodward, A. L. (2003). Infants' developing understanding of the link between looker and object. *Developmental Science, 6*(3), 297–311.
Yoshida, H., & Smith, L. B. (2008). What's in view for toddlers? Using a head camera to study visual experience. *Infancy, 13*(3), 229–248.
Yu, C., & Smith, L. B. (2012). Embodied attention and word learning by toddlers. *Cognition, 125*, 244–262.
Yu, C., & Smith, L. B. (2013). Joint attention without gaze following: Human infants and their parents coordinate visual attention to objects through eye-hand coordination. *PLoS One, 8*(11), e79659. http://dx.doi.org/10.1371/journal.pone.0079659.
Yurovsky, D., Smith, L. B., & Yu, C. (2013). Statistical word learning at scale: The baby's view is better. *Developmental Science, 16*(6), 959–966.