¹ A communicative framework for early word learning

² Benjamin C. Morris[1] & Daniel Yurovsky[1,2]

³ ¹ University of Chicago

⁴ ² Carnegie Mellon University

⁵ Author Note

⁶ Correspondence concerning this article should be addressed to Benjamin C. Morris,

⁷ Department of Psychology, University of Chicago, 5848 S University Ave, Chicago, IL 60637.

⁸ E-mail: benmorris@uchicago.edu

9   Abstract

10   Children do not learn language from passive observation of the world, but from interaction

11   with caregivers who want to communicate with them. These communicative exchanges are

12   structured at multiple levels in ways that support support language learning. We argue this

13   pedagogically supportive structure can result from pressure to communicate successfully with

14   a linguistically immature partner. We first characterize one kind of pedagogically supportive

15   structure in a corpus analysis: caregivers provide more information-rich referential

16   communication, using both gesture and speech to refer to a single object, when that object is

17   rare and when their child is young. Then, in an iterated reference game experiment on

18   Mechanical Turk (n = 480), we show how this behavior can arise from pressure to

19   communicate successfully with a less knowledgeable partner. Lastly, we show that speaker

20   behavior in our experiment can be explained by a rational planning model, without any

21   explicit teaching goal. We suggest that caregivers' desire to communicate successfully may

22   play a powerful role in structuring children's input in order to support language learning.

23   *Keywords:* language learning; communication; computational modeling

24   Word count: X

₂₅   A communicative framework for early word learning

₂₆   One of the most striking aspects of children's language learning is just how quickly

₂₇ they master the complex system of their natural language (Bloom, 2000). In just a few short

₂₈ years, children go from complete ignorance to conversational fluency in a way that is the

₂₉ envy of second-language learners attempting the same feat later in life (Newport, 1990).

₃₀ What accounts for this remarkable transition?

₃₁   Distributional learning presents a unifying account of early language learning: Infants

₃₂ come to language acquisition with a powerful ability to learn the latent structure of language

₃₃ from the statistical properties of speech in their ambient environment (Saffran, 2003).

₃₄ Distributional learning mechanisms can be seen in accounts across language including

₃₅ phonemic discrminitation (Maye, Werker, & Gerken, 2002), word segmentation (Saffran,

₃₆ 2003), learning the meanings of both nouns (Smith & Yu, 2008) and verbs (Scott & Fisher,

₃₇ 2012), learning the meanings of words at multiple semantic levels (Xu & Tenenbaum, 2007),

₃₈ and perhaps even the grammatical categories to which a word belongs (Mintz, 2003). A

₃₉ number of experiments clearly demonstrate both the early availability of distributional

₄₀ learning mechanisms and their potential utility across these diverse language phenomena

₄₁ (DeCasper & Fifer, 1980; DeCasper & Spence, 1986; Gomez & Gerken, 1999; Graf Estes,

₄₂ Evans, Alibali, & Saffran, 2007; Maye, Werker, & Gerken, 2002; Saffran, Newport, & Aslin,

₄₃ 1996; Smith & Yu, 2008; Xu & Tenenbaum, 2007).

₄₄   However, there is reason to be suspicious about just how precocious statistical learning

₄₅ abilities are in early development. Although these abilities are available early, they are

₄₆ highly constrained by limits on other developing cognitive capacities. For example, infants'

₄₇ ability to track the co-occurrence information connecting words to their referents is

₄₈ constrained significantly by their developing memory and attention systems (Smith & Yu,

₄₉ 2013; Vlach & Johnson, 2013). Computational models of these processes show that the rate

₅₀ of acquisition is highly sensitive to variation in environmental statistics (e.g., Vogt, 2012).

Models of cross-situational learning have demonstrated that the Zipfian distribution of word frequencies and word meanings yields a learning problem that cross-situational learning alone cannot explain over a reasonable time frame (Vogt, 2012). Further, a great deal of empirical work demonstrates that cross-situational learning even in adults drops off rapidly when participants are asked to track more referents, and also when the number of intervening trials is increased (e.g., Yurovsky & Frank, 2015). Thus, precocious unsupervised statistical learning appears to fall short of a complete explanation for rapid early language learning.

Even relatively constrained statistical learning could be rescued, however, if caregivers structured their language in a way that simplified the learning problem and promoted learning. For example, in phoneme learning, infant-directed speech provides examples that seem to facilitate the acquisition of phonemic categories (Eaves et al., 2016). In word segmentation tasks, infant-directed speech facilitates infant learning more than matched adult-directed speech (Thiessen, Hill, & Saffran, 2005). In word learning scenarios, caregivers produce more speech during episodes of joint attention with young infants, which uniquely predicts later vocabulary (Tomasello & Farrar, 1986). Child-directed speech even seems to support learning at multiple levels in parallel– e.g., simultaneous speech segmentation and word learning (Yurovsky et al., 2012). For each of these language problems faced by the developing learner, caregiver speech exhibits structure that seems uniquely beneficial for learning.

Under distributional learning accounts, the existence of this kind of structure is a theory-external feature of the world that does not have an independently motivated explanation. Such accounts view the generative process of structure in the language environment as a problem separate from language learning. However, across a number of language phenomena, the language environment is not merely supportive, but seems calibrated to children's changing learning mechanisms. For example, across development, caregivers engage in more multimodal naming of novel objects than familiar objects, and rely

on this synchrony most with young children (Gogate, Bahrick, & Watson, 2000). The role of synchrony in child-directed speech parallels infant learning mechanisms: young infants appear to rely more on synchrony as a cue for word learning than older infants, and language input mirrors this developmental shift (Gogate, Bahrick, & Watson, 2000). Beyond age-related changes, caregiver speech may also support learning through more local calibration to a child's knowledge; caregivers have been shown to provide more language to refer to referents that are unknown to their child, and show sensitivity to the knowledge their child displays during a referential communication game (Leung et al., 2019). The calibration of parents production to the child's learning suggests a co-evolution such that these processes should not be considered in isolation.

What then gives rise to structure in early language input that mirrors child learning mechanisms? Because of widespread agreement that parental speech is not usually motivated by explicit pedagogical goals (Newport et al., 1977), the calibration of speech to learning mechanisms seems a happy accident; parental speech just happens to be calibrated to children's learning needs. Indeed, if parental speech was pedagogically-motivated, we would have a framework for deriving predictions and expectations (e.g., Shafto, Goodman, & Griffiths, 2014). Models of optimal teaching have been successfully generalized to phenomena as broad as phoneme discrimination (Eaves et al., 2016) to active learning (Yang et al., 2019). These models take the goal to be to teach some concept to a learner and attempt to optimize that learner's outcomes. While these optimal pedagogy accounts have proven impressively useful, such models are theoretically unsuited to explaining parent language production where there is widespread agreement that caregiver goals are not pedagogical (e.g., Newport et al., 1977).

Instead, the recent outpouring of work exploring optimal communication (the Rational Speech Act model, see Frank & Goodman, 2012) provides another framework for understanding parent production. Under optimal communication accounts, speakers and

103 listeners engage in recursive reasoning to produce and interpret speech cues by making

104 inferences over one another's intentions (Frank & Goodman, 2012). These accounts have

105 made room for advances in our understanding of a range of language phenomena previously

106 uncaptured by formal modeling, notably a range of pragmatic inferences (e.g., Frank &

107 Goodman, 2012; other RSA papers). In this work, we consider the communicative structure

108 that emerges from an optimal communication system across a series of interactions where

109 one partner has immature linguistic knowledge. This perspective offers the first steps toward

110 a unifying account of both the child's learning and the parents' production: Both are driven

111 by a pressure to communicate successfully (Brown, 1977).

112   Early, influential functionalist accounts of language learning focused on the importance

113 of communicative goals (e.g., Brown, 1977). Our goal in this work is to formalize the

114 intuitions in these accounts in a computational model, and to test this model against

115 experimental data. We take as the caregiver's goal the desire to communicate with the child,

116 not about language itself, but instead about the world in front of them. To succeed, the

117 caregiver must produce the kinds of communicative signals that the child can understand and

118 respond contingently, potentially leading caregivers to tune the complexity of their speech as

119 a byproduct of in-the-moment pressure to communicate successfully (Yurovsky, 2017).

120   To examine this hypothesis, we focus on ostensive labeling (i.e. using both gesture and

121 speech in the same referential expression) as a case-study phenomenon of information-rich

122 structure in the language learning environment. We first analyze naturalistic parent

123 communicative behavior in a longitudinal corpus of parent-child interaction in the home

124 (Goldin-Meadow et al., 2014). We investigate the extent to which parents tune their

125 ostensive labeling across their child's development to align to their child's developing

126 linguistic knowledge (Yurovsky, Doyle, & Frank, 2016).

127   We then experimentally induce this form of structured language input in a simple

128 model system: an iterated reference game in which two players earn points for

communicating successfully with each other. Modeled after our corpus data, participants are asked to make choices about which communicative strategy to use (akin to modality choice). In an experiment on Mechanical Turk using this model system, we show that tuned, structured language input can arise from a pressure to communicate. We then show that participants' behavior in our game conforms to a model of communication as rational planning: People seek to maximize their communicative success while minimizing their communicative cost over expected future interactions. Lastly, we demonstrate potential benefits for the learner through a series of simulations to show that communicative pressure facilitates learning compared with various distributional learning accounts.

Corpus Analysis

We first investigate parent referential communication in a longitudinal corpus of parent-child interaction. We analyze the production of multi-modal cues (i.e. using both gesture and speech) to refer to the same object, in the same instance. While many aspects of CDS support learning, multi-modal cues (e.g., speaking while pointing or looking) are particularly powerful sources of data for young children (e.g., Baldwin, 2000, @gogate2000). We take multi-modal cues to be a case-study pheonmenon of pedagogically supportive language input. While our account should hold for other language phenomena, by focusing on one phenomenon we attempt to specify the dynamics involved in the production of such input.

In this analysis of naturalistic communication, we examine the prevelance of multi-modal cues in children's language environment, to demonstrate that it is a viable, pedagogically supportive form of input. Beyond being a prevelant fform of communication, multi-modal reference may be especially pedagogically supportive if usage patterns reflect adaptive linguistic tuning, with caregivers using this information-rich cue more for young children and infrequent objects. The amount of multi-modal reference should be sensitive to the child's age, such that caregivers will be more likely to provide richer communicative

information when their child is younger (and has less linguistic knowledge) than as she gets older (Yurovsky, Doyle, & Frank, 2016).

Methods

We used data from the Language Development Project– a large-scale, longitudinal corpus of naturalistic parent child-interaction in the home (Goldin-Meadow et al., 2014). The Language Development Project corpus contains transcription of all speech and communicative gestures produced by children and their caregivers over the course of the 90-minute home recordings. An independent coder analyzed each of these communicative instances and identified each time a concrete noun was referenced using speech, gesture, or both in the same referential expression (so called ostenstive labeling). In these analyses, we focus only on caregiver's productions of ostenstive labeling.

Participants.   The Language Development Project aimed to recruit a sample of families who are representative of the Chicago community in socio-economic and racial diversity (Goldin-Meadow et al., 2014). These data are drawn from a subsample of 10 families from the larger corpus. Our subsample contains data taken in the home every 4-months from when the child was 14-months-old until they were 34-months-old, resulting in 6 timepoints (missing one family at the 30-month timepoint). Recordings were 90 minute sessions, and participants were given no instructions.

Of the 10 target children, 5 were girls, 3 were Black and 2 were Mixed-Race. Families spanned a broad range of incomes, with 2 families earning $15,000 to $34,999 and 1 family earning greater than $100,000. The median family income was $50,000 to $74,999.

Procedure.   From the extant transcription and gesture coding, we specifically coded all concrete noun referents produced in either the spoken or gestural modality (or both). Spoken reference was coded only when a specific noun form was used (e.g., "ball"), to exlcude pronouns and anaphoric usages (e.g., "it"). Gesture reference was coded only for deitic gestures (e.g., pointing to or holding an object) to minimize ambiguity in determining

the intended referent. In order to fairly compare rates of communication across modalities, we need to examine concepts that can be referred to in either gesture or speech (or both) with similar ease. Because abstract entites are difficult to gesture about using deitic gestures, we coded only on references to concrete nouns.

Reliability.   To establish the reliability of the referent coding, 25% of the transcripts were double-coded. Inter-rater reliability was sufficently high (Cohen's $\kappa = 0.76$). Disagreements in coding decisions were discussed and resolved by hand.

To ensure that our each referent could potentially be refered to in gesture or speech, we focused on concrete nouns. We further wanted to ensure that the referents were physically present in the scene (and thus accessible to deitic gestures). Using the transcripts, a human rater judged whether the referent was likely to be present, primarily relying on discourse context (e.g., a referent was coded as present if the deitic gesture is used or used at another timepoint for the reference, or if the utterance included demonstratives such as "This is an X"). A full description of the coding criterea can be found in the linked coding manual.

To ensure our transcript-based coding of presentness was sufficiently accurate, a subset of the transcripts (5%) were directly compared to corresponding video data observation. Reliability across the video data and the transcript coding was sufficiently high ($\kappa = 0.72$). Based on transcript coding of all the referential communication about concrete nouns, 90% of the references were judged to be about referents that were likely present. All references are included in our dataset for further analysis.

Results

These corpus data were analyzed using a mixed effects regression to predict parent use of multi-modal reference for a given referent. The model included fixed effects of age in months, frequency of the referent, and the interaction between the two. The model included a random intercept and random slope of frequency by subject and a random intercept for

each unique referent. Frequency and age were both log-scaled and then centered both

because age and frequency tend to have log-linear effects and to help with model convergence.

The model showed that parents teach less to older children (β = -0.78, t = -7.88, $p <$ .001),

marginally less for more frequent targets (β = -0.08, t = -1.81, p = .071), and that parents

teach their younger children more often for equally frequent referents (β = 0.18, t = 3.25, p =

.001). Thus, in these data, we see early evidence that parents are providing richer, structured

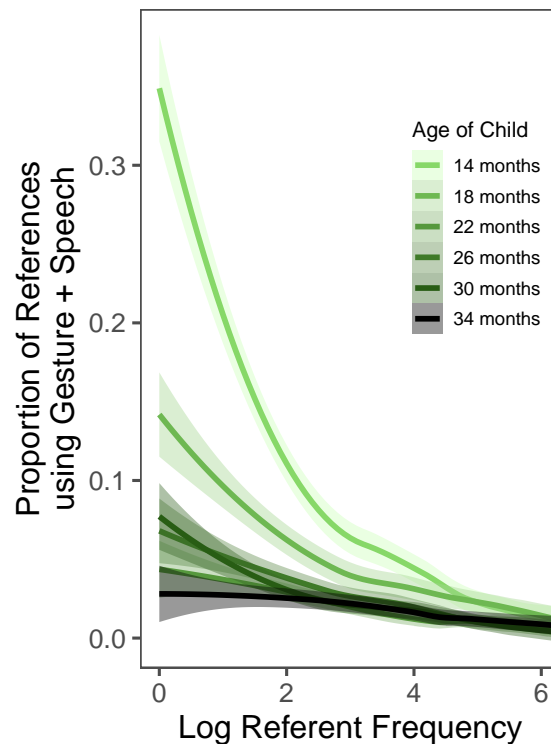input about rarer things in the world for their younger children (Figure \ref{fig:corpus-plot).

Figure 1. Proportion of parent multi-modal referential talk across development. The log of a

referent's frequency is given on the x-axis, with less frequent items closer to zero.

Discussion

Caregivers are not indiscriminate in their use of multi-modal reference; in these data,

they provided more of this support when their child was younger and when discussing less

familiar objects. These longitudinal corpus findings are consistent with an account of

parental alignment: parents are sensitive to their child's linguistic knowledge and adjust

their communication accordingly (Yurovsky et al., 2016). Ostensive labeling is perhaps the most explicit form of pedagogical support, so we chose to focus on it for our first case study. We argue that these data could be explained by a simple, potentially-selfish pressure: to communicate successfully. The influence of communicative pressure is difficult to draw in naturalistic data, so we developed a paradigm to try to experimentally induce richly-structured, aligned input from a pressure to communicate in the moment.

Experimental Framework

To study the emergence of pedagogically supportive input from communicative pressure, we developed a simple reference game in which participants would be motivated to communicate successfully. After giving people varying amounts of training on novel names for 9 novel objects, we asked them to play a communicative game in which they were given one of the objects as their referential goal, and they were rewarded if their partner successfully selected this referent from among the set of competitors (Figure **??**).

Participants could choose to refer either using the novel labels they had been exposed to, or they could use a deictic gesture to indicate the referent to their partner. Deixis was unambiguous, and thus would always succeed. However, in order for language to be effective, the participant and their partner would have to know the correct novel label for the referent.

Across conditions, we manipulated the relative costs of these two communicative methods (gesture and speech), as we did not have a direct way of assessing these costs in our naturalistic data, and they likely vary across communicative contexts. In all cases, we assumed that gesture was more costly than speech. Though this need not be the case for all gestures and contexts, our framework compares simple lexical labeling and unambiguous deictic gestures, which likely are more costly and slower to produce (see Yurovsky, 2018). We set the relative costs by explicitly implementing strategy utility, assigning point values to each communicative method.

²⁴³   If people are motivated to communicate successfully, their choice of referential modality

²⁴⁴ should reflect the tradeoff between the cost of producing the communicative signal with the

²⁴⁵ likelihood that the communication would succeed. We thus predicted that peoples' choice of

²⁴⁶ referential modality would reflect this tradeoff: People should be more likely to use language

²⁴⁷ if they have had more exposures to the novel object's correct label, and they should be more

²⁴⁸ likely to use language as gesture becomes relatively more costly.

²⁴⁹   Across two experiments, XXX participants were recruited to play our reference games

²⁵⁰ via Amazon Mechanical Turk, an online platform that allows workers to complete surveys

²⁵¹ and short tasks for payment. In these studies, all participants were placed in the role of

²⁵² speaker and listener responses were programmed.

Experiment 1

²⁵⁴   In this experiment, we provide proof of concept demonstrating that participants were

²⁵⁵ sensitive to our manipulations and that we could induce speech-gesture tradeoff with this

²⁵⁶ paradigm.

Method

²⁵⁸   **Participants.**   80 participants were recruited though Amazon Mechanical Turk and

²⁵⁹ received a small payment for their participation. Data from XXX participants was excluded

²⁶⁰ from subsequent analysis for failing the manipulation check or for producing illegal

²⁶¹ pseudo-English labels (e.g., "pricklyyone").

²⁶²   **Design and Procedure.**   Participants were told they would be introduced to novel

²⁶³ object-label pairs and then asked to play a communication game with a partner wherein they

²⁶⁴ would have to refer to a particular target object. Participants were exposed to nine novel

²⁶⁵ objects, each with a randomly assigned pseudo-word label. We manipulated the exposure

²⁶⁶ rate within-subjects: during training participants saw three of the nine object-label

²⁶⁷ mappings four times, two times, or just one time, yielding a total of 21 training trials.

268 Participants were then given a simple recall task to establish their knowledge of the novel

269 lexicon (pretest).

270   After being introduced to the rules of the game, participants are screened to ensure

271 they understand the rules of the game (manipulation check). During gameplay, speakers saw

272 the target object in addition to an array of all six objects. Speakers had the option of either

273 directly selecting the target object from the array (deictic gesture)- a higher cost cue but

274 without ambiguity- or typing a label for the object (speech)- a lower cost cue but contingent

275 on the listener's knowledge. After sending the message, speakers are shown which object the

276 listener selected.

277   If the speaker clicked on object (gesture message), the listener was programmed to

278 simply make the same selection. To simulate knowledgable listener behavior when the

279 speaker typed an object label in Experiment 1, the listener evaluated the Levenshtein

280 distance (LD) between the typed label and each of the nine possible labels and selected the

281 candidate with the smallest edit distance (e.g., if a speaker entered the message "tomi", the

282 programmed listener would select the referent corresponding to "toma"). If the speaker

283 message had an LD greater than two with each of the nine words in the novel lexicon, the

284 listener always selected an incorrect object.

285   Speakers could win up to 100 points per trial if the listener correctly selected the target

286 referent based on their message. If the listener failed to identify the target object, the

287 speaker received no points. We manipulated the relative utilities of each of the strategies

288 between-subjects. In the "Higher Relative Cost" condition, speakers received 30 points for

289 gesturing and 100 points for labeling, and thus gesturing is very costly relative to speech and

290 pariticpants should highly incentivized to speak. In the "Low Relative Cost" condition

291 speakers received 50 points for gesturing and 80 points for labeling, and thus gesturing is still

292 costly relative to speech but much less so and pariticpants should be less incentivized to

293 speak. 40 participants were run in each of the two conditions.

294   Speakers could win up to 100 points per trial if the listener correctly selected the target

295   referent. We manipulated the relative utility of the speech cue between-subjects across two

296   conditions: low relative cost for speech ("Low Relative Cost") and higher relative cost for

297   speech ("Higher Relative Cost"). In the "Low Relative Cost" condition, speakers were

298   charged 70 points for gesturing and 0 points for labeling, yielding 30 points and 100 points

299   respectively if the listener selected the target object. In the "Higher Relative Cost"

300   condition, speakers were charged 50 points for gesturing and 20 points for labeling, yielding

301   up to 50 points and 80 points respectively. If the listener failed to identify the target object,

302   the speaker nevertheless paid the relevant cost for that message in that condition. As a

303   result of this manipulation, there was a higher relative expected utility for labeling in the

304   "Low Relative Cost" condition than the "Higher Relative Cost" condition.

305   **Results.**

Experiment 2

307   Thus far, we have focused on relatively straightforward scenarios to demonstrate that a

308   pressure to communicate successfully in the moment can lead speakers to trade-off between

309   gesture and speech sensibly. However, critical to these repeated interactions is the ability to

310   learn about an interlocutor and potentially influence their learning. In Experiment 2,

311   participants were told about a third type of message: using both gesture and speech within a

312   single trial to effectively teach the listener an object-label mapping. This strategy

313   necessitates making inferences about the listener's knowledge state, so in Experiment 2 we

314   induced knowledge asymmetries between speaker and listner. As in Experiment 1,

315   pariticpants are trained and tested on 9 novel objects to establish their own knowledge, but

316   in Experiment 2 we also manipulated how much training they thought their partner received.

317   Using these manipulations, we aimed to experimentally determine the circumstances under

318   which richly-structured input emerges, without an explicit pedagogical goal.

Method

320   **Participants.**   480 participants were recruited though Amazon Mechanical Turk and

321 received $1 for their participation. Data from 51 participants were excluded from subsequent

322 analysis for failing the critical manipulation check and a further 28 for producing

323 pseudo-English labels (e.g., "pricklyyone"). The analyses reported exclude the data from

324 those participants, but all analyses were also conducted without excluding any participants

325 and all patterns hold ($ps < 0.05$).

326   **Design and Procedure.**   In Experiment 2, training and pretest were identical to

327 Experiment 1 protocols. Participants were again randomly assigned to one of two utlity

328 conditions: "Higher Relative Cost" and "Low Relative Cost." These conditions were matched

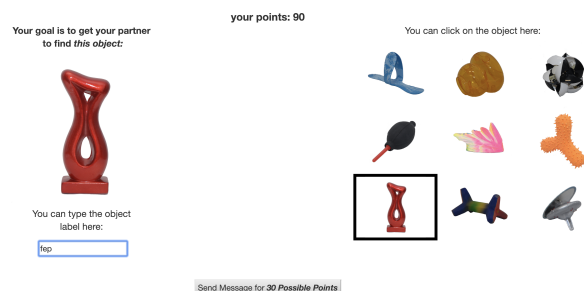329 to our Experiment 1 design, outlined above.

Figure 2. (#fig:exp_screenshot)Screenshot of speaker view during gameplay.

330   In Experiment 2, we also manipulated participants' expectations about their partner's

331 knowledge to explore the role of knowledge asymmetries. Prior to beginning the game,

332 participants were told how much exposure their partner had to the lexicon and also that they

333 would be asked to discuss each object three times. Across 3 between subjects conditions,

334 participants were told that their partner had either no experience with the lexicon, had the

335 same experience as the speaker, or had twice the experience of the speaker. As a

336 manipulation check, participants were then asked to report their partner's level of exposure,

337 and were corrected if they answer incorrectly. Gameplay then proceeded as in Experiment 1,

338 though listener behavior was programmed to match their supposed knowledge state.

Listeners were programmed with starting knowledge states initialized accordingly. Listeners with no exposure began the game with knowledge of 0 object-label pairs. Listeners with the same exposure of the speaker began with knowledge of five object-label pairs (3 high frequency, 1 mid frequency, 1 low frequency), based the average retention rates found previously. Lastly, the listener with twice as much exposure as the speaker began with knowledge of all nine object-label pairs. If the speaker produced a label, the listener was programmed to consult their own knowledge of the lexicon and check for similar labels (selecting a known label with a Levenshtein edit distance of two or fewer from the speaker's production), or select among unknown objects if no similar labels are found. Listeners could integrate new words into their knowledge of the lexicon if taught.

Critically, participants were told about a third type of possible message using both gesture and speech within a single trial to effectively teach the listener an object-label mapping. This action directly mirrors the multi-modal reference behavior from our corpus data– it presents the listener with an information-rich, potentially pedagogical learning moment. In order to produce this teaching behavior, speakers had to pay the cost of producing both cues (i.e. both gesture and speech). Note that, in all utility conditions, teaching yielded participants 30 points (compared with the much more beneficial strategy of speaking which yielded 100 points or 80 points across our two utility manipulations).

Crossing our 2 between-subjects manipulations yielded 6 conditions (2 utility manipulations: "Low Relative Cost" and "Higher Relative Cost"; and 3 levels of partner's exposure: None, Same, Double), with 80 participants in each condition. We expected to find results that mirrored our corpus findings such that rates of teaching would be higher when there was an asymmetry in knowledge where the speaker knew more (None manipulation) compared with when there was equal knowledge (Same manipulation) or when the listener was more familiar with the language (Double manipulation). We expected that participants would also be sensitive to our utility manipulation, such that rates of labeling and teaching

365 would be higher in the "Low Relative Cost" conditions than the other conditions.

Results

367   As an initial check of our exposure manipulation, we fist a logistic regression predicting

368 accuracy at test from a fixed effect of exposure rate and random intercepts and slopes of

369 exposureRate by participant as well as random intercepts by item. We found a reliable effect

370 of exposure rate, indicating that participants were better able to learn items that appear

371 more frequently in training ($\beta = 1.09$, $t = 13.73$, $p < .001$). On average, participants knew

372 at least 6 of the 9 words in the lexicon (mean $= 6.28$, sd $= 2.26$).
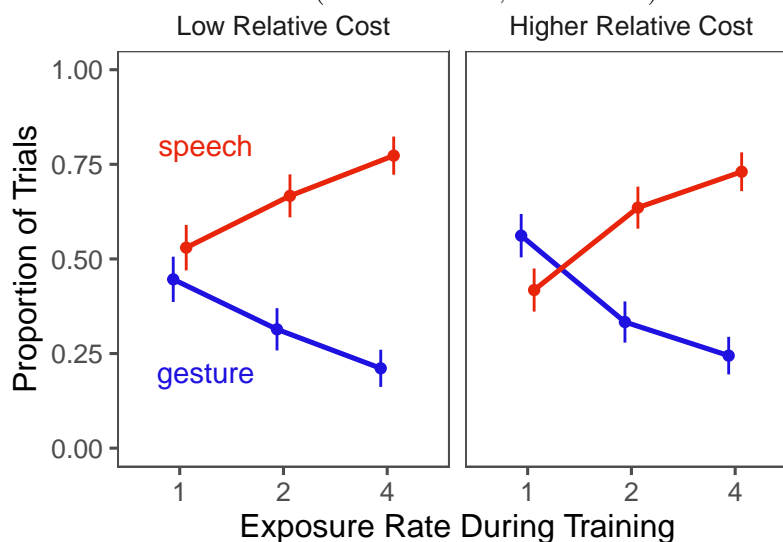

Figure 3. (#fig:speech_gesture)Speaker communicative method choice as a function of exposure and the utility manipulation.

373   **Gesture-Speech Tradeoff.**   Figure **??** illustrates the gesture-speech tradeoff

374 pattern in the Double Exposure condition (as there was minimal teaching in that condition,

375 so the speech-gesture trade-off is most interpretable). The effects on gesture mirror those

376 found for labeling and are thus not included for brevity (*ps < 0.01*). Note that these effects

377 cannot be explained by participant knowledge; all patterns above hold when looking *only* at

378 words known by the speaker at pretest (*ps < 0.01*). Further, these patterns directly mirror

379 previous corpus analyses demonstrating the gesture-speech tradeoff in naturalistic parental

communicative behaviors, where lexical knowledge is likely for even the least frequent referent (see Yurovsky, 2018).

Emergence of Teaching.   Thus far, we have focused on relatively straightforward scenarios to demonstrate that a pressure to communicate successfully in the moment can lead speakers to trade-off between gesture and speech sensibly. Next, we turn to the emergence of teaching behavior.

Discussion

As predicted, the data from our paradigm corroborate our findings from the corpus analysis, demonstrating that pedagogically supportive behavior emerges despite the initial cost when there is an asymmetry in knowledge and when speech is less costly than other modes of communication. While this paradigm has stripped away much of the interactive environment of the naturalistic corpus data, it provides important proof of concept that the structured and tuned language input we see in those data could arise from a pressure to communicate. The paradigm's clear, quantitative predictions also allow us to build a formal model to predict our empirical results.
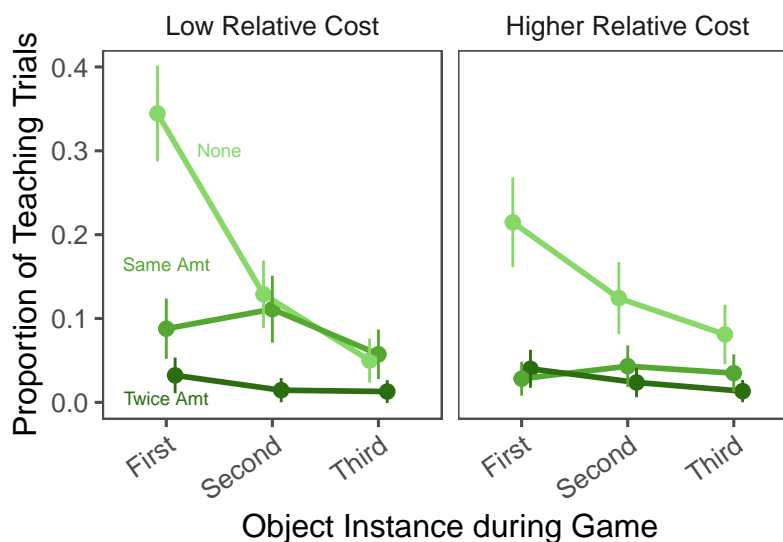

Figure 4. (#fig:exp_teach)Rates of teaching across the 6 conditions, plotted by how many times an object had been the target object.

The results from this experiment are qualitatively consistent with a model in which participants make their communicative choices to maximize their expected utility from the reference game. We next formalize this model to determine if these results are predicted quantitatively as well.

Model: Communication as planning

In order to model when people should speak, point, or teach, we begin from the problem of what goal people are trying to solve (Marr, 1982). Following a long history of work in philosophy of language, we take the goal of communication to be causing an action in the world by transmitting some piece of information to one's conversational partner (e.g. Wittgenstein, 1953; Austin, 1975). If people are near-optimal communicators, they should choose communicative signals that maximize the probability of being understood while minimizing the cost of producing the signal (Clark, 1996; Grice, 1975). In the special case of reference, solving this problem amounts to producing the least costly signal that correctly specifies one's intended target referent in such a way that one's conversational partner can select it from the set of alternative referents.

Recently, Frank and Goodman (2012) developed the Rational Speech Act framework–a formal instantiation of these ideas. In this model, speakers choose from a set of potential referential expressions in accordance to a utility function that maximizes the probability that a listener will correctly infer their intended meaning while minimizing the number of words produced. This framework has found successful application in a variety of linguistic applications such as scalar implicature, conventional pact formation, and production and interpretation of hyperbole (Goodman & Frank, 2016; see also related work from Franke, 2013). These models leverage recursive reasoning–speakers reasoning about listeners who are reasoning about speakers–in order to capture cases in which the literal meaning and the intended meaning of sentences diverge.

420   To date, this framework has been applied primarily in cases where both communicative

421 partners share the same linguistic repertoire, and thus communicators know their probability

422 of communicating successfully having chosen a particular signal. This is a reasonable

423 assumption for pairs of adults in contexts with shared common ground. But what if partners

424 do not share the same lingusitic repertoire, and in fact do not know the places where their

425 knowledge diverges? In this case, communicators must solve two problems jointly: (1) Figure

426 out what their communicative partner knows, and (2) produce the best communicative

427 signal they can given their estimates of their partner's knowledge. If communicative partners

428 interact repeatedly, these problems become deeply intertwined: Communicators can learn

429 about each-other's knowledge by observing whether their attempts to communicate succeed.

430 For instance, if a communicator produces a word that identifies their intended referent, but

431 their partner fails to select that referent from among the set of objects, they can infer that

432 their partner must not share their understanding of this word. They might then choose not

433 to use language to refer to this object in the future, but choose to point to it instead.

434   Critically, communicators can also change each-other's knowledge. When a

435 communicator both points to an object and produces a linguistic label, they are in effect

436 teaching their partner the word that they use to refer to this object. While this this behavior

437 is costly in the moment, and no more referentially effective than pointing alone, it can lead to

438 more efficient communication in the future–instead of pointing to this referent forever more,

439 communicators can now use the linguistic label they both know they share. This behavior

440 naturally emerges from a conception of communication as planning: Communicators' goal is

441 to choose a communicative signal today that will lead to efficient communication not just in

442 the present moment, but in future communications as well. If they are likely to need to refer

443 to this object frequently, it is worth it to be inefficient in this one exchange in order to be

444 more efficient future. In this way, pedagogically supportive behavior can emerge naturally

445 from a model with no explicit pedagogical goal. In the following section, we present a formal

446 instantiation of this intuitive description of communication as planning and show that it

accounts for the behavior we observed in our experiments.

Alternatively, pedogically-supportive input could emerge from an explicit pedagogical goal. Shafto, Goodman, and Griffiths (2014) have developed an framework of rational pedagogy built on the same recursive reasoning principles as in the Rational Speech Act Framework: Teachers aim to teach a concept by choosing a set of examples that would maximize learning for students who reason about the teachers choices as attempting to maximize their learning. Rafferty, Brunskill, Griffiths, and Shafto (2016) et al expanded framework to sequential teaching, in which teachers use students in order to infer what they have learned and choose the subsequent example. In this case, teaching can be seen as a kind of planning where teachers should choose a series of examples that will maximize students learning but can change plans if an example they thought would be too hard turns out too easy–or vice-versa. In the case of our reference game, this model is indistinguishable form a communicator seeks to maximize communicative success but is indifferent to communicative cost. This model makes poor predictions about parents' behavior in our corpus, and also adults' behavior in our experiments, but we return to it in the subsequent section to consider how differences in parents' goals and differences in children's learning contribute to changes in the rate of language acquisition.

Formal Model

We take as inspiration the idea that communication is a kind of action–e.g. talking is a speech act (Austin, 1975). Consequently, we can understand the choice of *which communicative act* a speaker should take as a question of which act would maximize their utility: achieving successful communication while minimizing their cost (Frank & Goodman, 2012). In this game, speakers can take three actions: talking, pointing, or teaching. In this reference game, these Utilities (U) are given directly by the rules. Because communication is a repeated game, people should take actions that maximize their Expected Utility (EU) over the course of not just this act, but all future communicative acts with the same

473 conversational partner. We can think of communication, then as a case of recursive planning.

474 However, people do not have perfect knowledge of each-other's vocabularies (v). Instead,

475 they only have uncertain beliefs (b) about these vocabularies that combine their expectations

476 about what kinds of words people with as much linguistic experience as their partner are

477 likely to know with their observations of their partner's behavior in past communicative

478 interactions. This makes communication a kind of planning under uncertainty well modeled

479 as a Partially Observable Markov Decision Process (POMDP, Kaelbling, Littman, &

480 Cassandra, 1998).

481   Optimal planning in a Partially Observable Markov Decision Process involves a cycle of

482 four phases: (1) Plan, (2) Act, (3) Observe, (4) Update beliefs. When people plan, they

483 compute the Expected Utility of each possible action (a) by combining the Expected Utility

484 of that action now with the Discounted Expected Utility they will get in all future actions.

485 The amount of discounting (γ) reflects how people care about success now compared to

486 success in the future. In our simulations, we set $\gamma = .5$ in line with prior work. Because

487 Utilities depend on the communicative partner's vocabulary, people should integrate over all

488 possible vocabularies in proportion to the probability that their belief assigns to that ($\mathbb{E}_{v \sim b}$).

$$EU\left[a\,|\,b\right] = \mathbb{E}_{v \sim b}\left(U(a|v) + \gamma\,\mathbb{E}_{v',o',a'}\left(EU\left[a'|b'\right]\right)\right)$$

489 Next, people take an action as a function of its Expected Utility. Following other models in

490 the Rational Speech Act framework, we use the Luce Choice Axiom, in which each choice is

491 taken in probability proportional to its exponentiated utility (Frank & Goodman, 2012; Luce,

492 1959). This choice rule has a single parameter α that controls the noise in this choice–as α

493 approaches 0, choice is random and as α approaches infinity, choice is optimal. For the results

494 reported here, we set $\alpha = 2$ based on hand-tuning, but other values produce similar results.

$$P\left(a|b\right) \propto \alpha\,e^{EU\left[a|b\right]}$$

495 After taking an action, people observe (o) their partner's choice–sometimes they pick

the intended object, and sometimes they do not. They then update their beliefs about the partner's vocabulary based on this observation. For simplicity, we assume that people think their partner should always select the correct target if they point to it, or if they teach, and similarly should always select the correct target if they produce its label and the label is in their partner's vocabulary. Otherwise, they assume that their partner will select the wrong object. People could of course have more complex inferential rules, e.g. assuming that if their partner does know a word they will choose among the set of objects whose labels they do not know (mutual exclusivity, Markman & Wachtel, 1988). Empirically, however, our simple model appears to accord well with people's behavior.

$$b'(v') \propto P(o|v', a) \sum_{v \in V} P(v'|v, a) \, b(v)$$

The critical feature of a repeated communication game is that people can change their partner's vocabulary. In teaching, people pay the cost of both talking and pointing together, but can leverage their partner's new knowledge on future trials. Note here that teaching has an upfront cost and the only benefit to be gained comes from using less costly communication modes later. There is no pedagogical goal– the model treats speakers as selfish agents aiming to maximize their own utilities by communicating successfully. We assume for simplicity that learning is approximated by a simple Binomial learning model. If someone encounters a word w in an unambiguous context (e.g. teaching), they add it to their vocabulary with probability p. We also assume that over the course of this short game that people do not forget–words that enter the vocabulary never leave, and that no learning happens by inference from mutual exclusivity.

$$P(v'|v, a) = \begin{cases} 1 & \text{if } v_w \in v \& v' \\ p & \text{if } v_w \notin v \& a = \text{point+talk} \\ 0 & otherwise \end{cases}$$

516   The final detail is to specify how people estimate their partner's learning rate (p) and

517 initial vocabulary (v). We propose that people begin by estimating their own learning rate

518 by reasoning about the words they learned at the start of the task: Their p is the rate that

519 maximizes the probability of them having learned their initial vocabularies from the trials

520 they observed. People can then expect their partner to have a similar p (per the "like me"

521 hypothesis, Meltzoff, 2005). Having an estimate of their partner's p, they can estimate their

522 vocabulary by simulating their learning from the amount of training we told them their

523 partner had before the start of the game.

Model Results

525   The fit between our model's predictions and our empirical data from our reference

526 game study on Amazon Turk can be seen in Figure **??**. The model outputs trial-level action

527 predictions (e.g., "speak") for every speaker in our empirical data. These model outputs

528 were aggregated across the same factors as the empirical data: modality, appearance,

529 partner's exposure, and utility condition. We see a significant correlation of our model

530 predictions and our empirical data ($r =$, $p<0.0001$). Our model provides a strong fit for

531 these data, supporting our conclusion that richly-structured language input could emerge

532 from in-the-moment pressure to communicate, without a goal to teach.

<center>**Consequences for Learning**</center>

534   In the model and experiments above, we asked whether the pressure to communicate

535 successfully with a linguistically-naive partner would lead to pedagogically supportive input.

536 These results confirmed its' sufficiency: As long as linguistic communication is less costly

537 than deictic gesture, speakers should be motivated to teach in order to reduce future

538 communicative costs. Further, the strength of this motivation is modulated by predictable

539 factors (speaker's linguistic knowledge, listener's linguistic knowledge, relative cost of speech

540 and gesture, learning rate, etc.), and the strength of this modulation is well predicted by a

541 rational model of planning under uncertainty about listner's vocabulary.

In this final section, we take up the consequences of communicatively-motivated teaching for the listener. To do this, we adapt a framework used by Blythe, Smith, and Smith (2010) and colleagues to estimate the learning times for an idealized child learning language under a variety of models of both the child and their parent. We come to these estimates by simulating exposure to successive communicative events, and measuring the probability that successful learning happens after each event. The question of how different models of the parent impact the learner can then be formalized as a question of how much more quickly learning happens in the context of one model than another.

We consider three parent models:

1. *Teacher* - under this model, we take the parents' goal to be maximizing the child's linguistic development. Each communicative event in this model consists of an ostensive labelling event (Note: this model is equivalent to a *Communicator* that ignores communicative cost).

2. *Communicator* - under this model, we take the parents' goal to be maximizing communicative success while minimizing communicative cost. This is the model we explored in the previous section.

3. *Indifferent* - under this model, the parent produces a linguistic label in each communicative event regardless of the child's vocabulary state. (Note: this model is equivalent to a *Communicator* who ignores communicative success).

SOME STUFF ABOUT CROSS SITUATIONAL LEARNING

One important point to note is that we are modeling the learning of a single word rather than the entirety of a multi-word lexicon (as in Blythe et al., 2010). Although learning times for each word could be independent, an important feature of many models of word learning is that they are not (Frank, Goodman, & Tenenbaum, 2009; Yu, 2008;

566 Yurovsky, Fricker, Yu, & Smith, 2014; although c.f. McMurray, 2007). Indeed, positive

567 synergies across words are predicted by the majority of models and the impact of these

568 synergies can be quite large under some assumptions about the frequency with which

569 different words are encountered (Reisenauer, Smith, & Blythe, 2013). We assume

570 independence primarily for pragmatic reasons here–it makes the simulations significantly

571 more tractable (although it is what our experimental participants appear to assume about

572 learners). Nonetheless, it is an important issue for future consideration. Of course, synergies

573 that support learning under a cross-situational scheme must also support learning from

574 communcators and teachers (Frank et al., 2009; Markman & Wachtel, 1988; Yurovsky, Yu, &

575 Smith, 2013). Thus, the ordering across conditions should remain unchanged. However, the

576 magnitude of the difference sacross teacher conditions could potentially increase or decrease.

Method

578 **Teaching.** Because the teaching model is indifferent to communicative cost, it

579 engages in ostensive an ostensive labeling (pointing + speaking) on each communicative

580 event. Consequently, learning on each trial occurs with a probability that depends entirely

581 on the learner's learning rate ($P_k = p$). Because we do not allow forgetting, the probability

582 that a learner has failed to successfully learn after n trials is equal to the probability that

583 they have failed to learn on each of n successive independent trials (The probabiliy of zero

584 successess on n trials of a Binomial random variable with parameter p). The probability of

585 learning after n trials is thus:

$$P_k(n) = 1 - (1 - p)^n$$

586 The expected probability of learning after n trials was thus defined analytically and

587 required no simulation. For comparison to the other models, we computed P_k for values of p

588 that ranged from .1 to 1 in increments of .1.

589   **Communication.**   To test learner under the communication model, we implemented

590 the same model described in the paper above. However, because our interest was in

591 understanding the relationship between parameter values and learning outcomes rather than

592 inferring the parameters that best describe people's behavior, we made a few simplifying

593 assumptions to allow many runs of the model to complete in a more practical amount of

594 time. First, in the full model above, speakers begin by inferring their own learning

595 parameters (P_s) from their observations of their own learning, and subsequently use their

596 maximum likelihood estimate as a standin for their listener's learning parameter (P_l).

597 Because this estimate will converge to the true value in expectation, we omit these steps and

598 simply stipulate that the speaker correctly estimates the listener's learning parameter.

599   Second, unless the speaker knows apriori how many times they will need to refer to a

600 particular referent, the planning process is an infinite recursion. However, each future step in

601 the plan is less impactful than the previous step (because of exponential discounting), this

602 infinite process is in practice well approximated by a relatively small number of recursive

603 steps. In our explorations we found that predictions made from models which planned over 3

604 future events were indistinguishable from models that planned over four or more, so we

605 simulated 3 steps of recursion[1]. Finally, to increase the speed of the simulations we

606 re-implemented them in the R programming language. All other aspects of the model were

607 identical.

608   **Hypothesis Testing.**   The literature on cross-situational learning is rich with a

609 variety of models that could broadly be considered to be "hypothesis testers." In an

610 eliminative hypothesis testing model, the learner begins with all possible mappings between

611 words and objects and prunes potential mappings when they are inconsistent with the data

612 according to some principe. A maximal version of this model relies on the principle that

---

[1] It is an intersting empirical question to determine how the level of depth to which that people plan in this
and similar games (see e.g. bounded rationality in Simon, 1991; resource-rationality in Griffiths, Lieder, &
Goodman, 2015). This future work is outside the scope of the current project.

every time a word is heard its referent must be present, and thus prunes any word-object

mappings that do not appear on the current trial. This model converges when only one

hypothesis remains and is probably the fastest learner when its assumed principle is a correct

assumption (Smith, Smith, & Blythe, 2011).

A positive hypothesis tester begins with no hypotheses, and on each trial stores one ore

more hypotheses that are consistent with the data, or alternatively strengthens one or more

hypotheses that it has already stored that are consistent with the new data. A number of

such models have appeared in the literature, with different assumptions about (1) how many

hypotheses a learner can store, (2) existing hypotheses are strengthened, (3) how existing

hypotheses are pruned, and (4) when the model converges (Siskind, 1996; Smith et al., 2011;

Stevens, Gleitman, Trueswell, & Yang, 2017; Trueswell, Medina, Hafri, & Gleitman, 2013; Yu

& Smith, 2012).

Finally, Bayesian models have been proposed that leverage some of the strengths of

both of these different kinds of model, both increasing their confidence in hypotheses

consisten with the data on a given learning event and decreasing their confidence in

hypotheses inconsistent with the event (Frank et al., 2009).

Because of its more natural alignment with the learning models we use Teaching and

Communication simulations, we implemented a positive hypothesis testing model[2]. In this

model, learners begin with no hypotheses and add new ones to their store as they encounter

data. Upon first encountering a word and a set of objects, the model encodes up to h

hypothesized word-object pairs each with probability p. On subsequent trials, the model

checks whether any of the existing hypotheses are consistent with the current data, and

---

[2] Our choice to focus on hypothesis testing rather than other learning frameworks is purely a pragmatic

choice–the learning parameter p in this models maps cleanly onto the learning parameter in our other models.

We encourage other researchers to adapt the code we have provided to estimate the long-term learning for

other models.

635 prunes any that are not. If no current hypotheses are consistent, it adds up to h new

636 hypotheses each with probability p. The model has converged when it has pruned all but the

637 one correct hypothesis for the meaning of a word. This model is most similar to the Propose

638 but Verify model proposed in Trueswell et al. (2013), with the exception that it allows for

639 multiple hypotheses. Because of the data generating process, storing prior disconfirmed

640 hypotheses (as in Stevens et al., 2017), or incrementing hypotheses consistent with some but

641 not all of the data (as in Yu & Smith, 2012) has no impact on learner and so we do not

642 implement it here. We note also that, as described in Yu and Smith (2012), hypothesis

643 testing models can mimic the behavior of associative learning models given the right

644 parameter settings (Townsend, 1990).

645   In contrast to the Teaching and Communication simulations, the behavior of the

646 Hypothesis Testing model depends on which particular non-target objects are present on

647 each naming event. We thus began each simulation by generating a copus of 100 naming

648 events, on each sampling the correct target as well as (C-1) competitors from a total set of

649 M objects. We then simulated a hypothesis tester learning over this set of events as

650 described above, and recorded the first trial on which the learner converged (having only the

651 single correct hypothesized mapping between the target word and target object). We

652 repeated this process 1000 times for each simulated combination of $M = (16, 32, 64, 128)$

653 total objects, $C = (1, 2, 4, 8)$ objects per trial, $h = (1, 2, 3, 4)$ concurrent hypotheses, as the

654 learning rate p varied from .1 to 1 in increments of .1.

General Discussion

656   Across naturalistic corpus data, experimental data, and model predictions, we see

657 evidence that pressure to communicate successfully with a linguistically immature partner

658 could fundamentally structure parent production. In our experiment, we showed that people

659 tune their communicative choices to varying cost and reward structures, and also critically to

660 their partner's linguistic knowledge–providing richer cues when partners are unlikely to know

661 the language and many more rounds remain. These data are consistent with the patterns

662 shown in our corpus analysis of parent referential communication and demonstrate that such

663 pedagogically supportive input could arise from a motivation to maximize communicative

664 success while minimizing communicative cost– no additional motivation to teach is necessary.

665 In simulation, we demonstrate that such structure could have profound implications for child

666 language learning, simplifying the learning problem posed by most distributional accounts of

667 language learning.

668   Accounts of language learning often aim to explain its striking speed in light of the

669 sheer complexity of the language learning problem itself. Many such accounts argue that

670 simple (associative) learning mechanisms alone seem insufficient to explain the rapid growth

671 of language skills and appeal instead to additional explanatory factors, such as the so-called

672 language acquisition device, working memory limitations, word learning biases, etc. (e.g.,

673 Chomsky, 1965; Goldowsky & Newport, 1993; Markman, 1990). While some have argued for

674 the simplifying role of language distributions (e.g., McMurray, 2007), these accounts largely

675 focus on learner-internal explanations. For example, Elman (1993) simulates language

676 learning under two possible explanations to intractability of the language learning problem:

677 one environmental, and one internal. He first demonstrates that learning is significantly

678 improved if the language input data is given incrementally, rather than all-at-once (Elman,

679 1993). He then demonstrates that similar benefits can arise from learning under limited

680 working memory, consistent with the "less-is-more" proposal (Elman, 1993; Goldowsky &

681 Newport, 1993). Elman dismisses the first account arguing that ordered input is implausible,

682 while shifts in cognitive maturation are well-documented in the learner (Elman, 1993);

683 however, our account's emphasis on changing calibration to such learning mechanisms

684 suggests the role of ordered or incremental input from the environment may be crucial.

685   This account is consonant with work in other areas of development, such as recent

686 demonstrations that the infant's visual learning environment has surprising consistency and

incrementality, which could be a powerful tool for visual learning. Notably, research using head mounted cameras has found that infant's visual perspective privileges certain scenes and that these scenes change across development (Fausey, Jayaraman, & Smith, 2016). In early infancy, the child's egocentric visual environment is dominated by faces, but shifts across infancy to become more hand and hand-object oriented in later infancy (Fausey et al., 2016). This observed shift in environmental statistics mirrors learning problems solved by infants at those ages, namely face recognition and object-related goal attribution respectively (Fausey et al., 2016). These changing environmental statistics have clear implications for learning and demonstrate that the environment itself is a key element to be captured by formal efforts to evaluate statistical learning (Smith et al., 2018). Frameworks of visual learning must incorporate both the relevant learning abilities and this motivated, contingent structure in the environment (Smith et al., 2018).

By analogy, the work we have presented here aims to draw a similar argument for the language environment, which is also demonstrably beneficial for learning and changes across development. In the case of language, the contingencies between learner and environment are even clearer than visual learning. Functional pressures to communicate and be understood make successful caregiver speech highly dependent on the learner. Any structure in the language environment that is continually suited to changing learning mechanisms must come in large part from caregivers themselves. Thus, a comprehensive account of language learning that can successfully grapple with the infant curriculum (Smith et al., 2018) must explain parent production, as well as learning itself. In this work, we have taken first steps toward providing such an account.

Explaining parental modification is a necessary condition for building a complete theory of language learning, but modification is certainly not a sufficient condition for language learning. No matter how callibrated the language input, non-human primates are unable to acquire language. Indeed, parental modification need not even be a necessary condition for

713 language learning. Young children are able to learn novel words from (unmodified) overheard

714 speech between adults (Foushee & Xu, 2016), although there is reason to think that

715 overheard sources may have limited impact on language learning broadly (e.g., Schniedman

716 & Goldin-Meadow, 2012). Our argument is that the rate and ultimate attainment of

717 language learners will vary substantially as a function of parental modification, and that

718 describing the cause of this variability is a necessary feature of models of language learning.

719 **Generalizability and Limitations.** Our account aims to think about parent

720 production and child learning in the same system, putting these processes into explicit

721 dialogue. While we have focused on ostensive labeling as a case-study phenomenon, our

722 account should reasonably extend to the changing structure found in other aspects of

723 child-directed speech– though see below for important limitations to this extension. Some

724 such phenomena will be easily accounted for: aspects of language that shape communicative

725 efficiency should shift in predictable patterns across development.

726 While these language phenomena can be captured by our proposed framework,

727 incorporating them will likely require altering aspects of our account and decisions about

728 which alterations are most appropriate. For example, the exaggerated pitch contours seen in

729 infant-directed speech could be explained by our account if we expand the definition of

730 communicative success to include a goal like maintaining attention. Alternatively, one could

731 likely accomplish the same goal by altering the cost and utility structure to penalize loss of

732 engagement. Thus, while this account should generalize to other modifications found in

733 child-directed speech, such generalizations will likely require non-trivial alterations to the

734 extant structure of the framework.

735 Of course, not all aspects of language should be calibrated to the child's language

736 development. Our account also provides an initial framework for explaining aspects of

737 communication that would not be modified in child-directed speech: namely, aspects of

738 communication that minimally effect communicative efficiency. In other words,

739 communication goals and learning goals are not always aligned. For example, children

740 frequently overregularize past and plural forms, producing incorrect forms such as "runn-ed"

741 (rather than the irregular verb "ran") or "foots" (rather than the irregular plural "feet")

742 (citation on overregularization). Mastering the proper tense endings (i.e. the learning goal)

743 might be aided by feedback from parent; however, adults rarely provide corrective feedback

744 for these errors (citation for lack of correction), perhaps because incorrect grammatical forms

745 are often sufficient to allow for successful communication (i.e. the communicative goal). The

746 degree of alignment between communication and learning goals should predict the extent to

747 which a linguistic phenomenon is modified in child-directed speech. Fully establishing the

748 degree to which modification is expected for a given language phenomena will likely require

749 working through a number of limitations in the generalizability of the framework as it stands.

750   Some aspects of parent production are likely entirely unrepresented in our framework,

751 such as aspects of production driven by speaker-side constraints. Furthermore, our account is

752 formulated primarily around concrete noun learning and future work must address its

753 viability in other language learning problems. We chose to focus on ostensive labeling as a

754 case-study phenomenon because it is an undeniably information-rich cue for young language

755 learners, however ostensive labeling varies substantially across socio-economic status and

756 cross-linguistically (citation for SES + lang ostensive labeling). This is to be expected to the

757 extent that parent-child interaction is driven by different goals (or goals given different

758 weights) across these populations– variability in goals could give rise to variability in the

759 degree of modification. Nonetheless, the generalizability of our account across populations

760 remains unknown. Indeed, child-directed speech itself varies cross-linguistically, both in its

761 features (citation) and quantity (citation). There is some evidence that CDS predicts

762 learning even in cultures where CDS is qualitatively different and less prevalent than in

763 American samples (Schneidman & Goldin-Meadow, 2012). Future work is needed to

764 establish the generalizability of our account beyond the western samples studied here.

765 We see this account as building on established, crucial statistical learning skills–

766 distributional information writ large and (unmodified) language data from overheard speech

767 are undoubtedly helpful for some learning problems (e.g., phoneme learning). There is likely

768 large variability in the extent to which statistical learning skills drive the learning for a given

769 learning problem. The current framework is limited by its inability to account for such

770 differences across learning problems, which could derive from domain or cultural differences.

771 Understanding generalizability of this sort and the limits of statistical learning will likely

772 require a full account spanning both parent production and child learning.

773 A full account that explains variability in modification across aspects of language will

774 rely on a fully specified model of optimal communication. Such a model will allow us to

775 determine both which structures are predictably unmodified, and which structures must be

776 modified for other reasons. Nonetheless, this work is an important first step in validating the

777 hypothesis that language input that is structured to support language learning could arise

778 from a single unifying goal: The desire to communicate effectively.

Conclusion

Acknowledgement

References

Austin, J. L. (1975). *How to do things with words* (Vol. 88). Oxford university press.

Baldwin, D. (2000). Interpersonal understanding fuels knowledge acquisition. *Current Directions in Psychological Science*, *9*, 40–45.

Blythe, R. A., Smith, K., & Smith, A. D. M. (2010). Learning times for large lexicons through cross-situational learning. *Cognitive Science*, *34*, 620–642.

Clark, H. H. (1996). *Using language. Journal of Linguistics* (pp. 167–222). Cambridge Univ Press.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*, 998–998.

Frank, M. C., Goodman, N., & Tenenbaum, J. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, *20*, 578–585.

Franke, M. (2013). Game theoretic pragmatics. *Philosophy Compass*, *8*(3), 269–284.

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. W., & Small, S. L. (2014). New evidence about language and cognitive development based on a longitudinal study: Hypotheses for intervention. *American Psychologist*, *69*(6), 588–599.

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, *20*(11), 818–829.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics: Vol. 3: Speech acts* (pp. 41–58). New York, NY: Academic Press.

Griffiths, T. L., Lieder, F., & Goodman, N. D. (2015). Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in Cognitive Science, 7*(2), 217–229.

Kaelbling, L. P., Littman, M. L., & Cassandra, A. R. (1998). Planning and acting in partially observable stochastic domains. *Artificial Intelligence, 101*, 99–134.

Luce, R. D. (1959). Individual choice behavior.

Markman, E. M., & Wachtel, G. F. (1988). Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive Psychology, 20*(2), 121–157.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information.* New York, NY: W. H. Freeman.

McMurray, B. (2007). Defusing the childhood vocabulary explosion. *Science, 317*(5838), 631–631.

Meltzoff, A. N. (2005). Imitation and other minds: The "like me" hypothesis. *Perspectives on Imitation: From Neuroscience to Social Science, 2*, 55–77.

Rafferty, A. N., Brunskill, E., Griffiths, T. L., & Shafto, P. (2016). Faster teaching via pomdp planning. *Cognitive Science, 40*(6), 1290–1332.

Reisenauer, R., Smith, K., & Blythe, R. A. (2013). Stochastic dynamics of lexicon learning in an uncertain and nonuniform world. *Physical Review Letters, 110*(25), 258701.

Shafto, P., Goodman, N. D., & Griffiths, T. L. (2014). A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive Psychology, 71*,

55–89.

Simon, H. A. (1991). Bounded rationality and organizational learning. *Organization Science*, *2*(1), 125–134.

Siskind, J. M. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, *61*, 39–91.

Smith, K., Smith, A. D., & Blythe, R. A. (2011). Cross-situational learning: An experimental study of word-learning mechanisms. *Cognitive Science*, *35*(3), 480–498.

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

Townsend, J. T. (1990). Serial vs. Parallel processing: Sometimes they look like tweedledum and tweedledee but they can (and should) be distinguished. *Psychological Science*, *1*(1), 46–54.

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.

Wittgenstein, L. (1953). *Philosophical investigations*. Oxford, UK: Basil Blackwell.

Yu, C. (2008). A statistical associative account of vocabulary growth in early word learning. *Language Learning and Development*, *4*(1), 32–62.

Yu, C., & Smith, L. B. (2012). Modeling cross-situational word-referent learning: Prior questions. *Psychological Review*, *119*, 21–39.

Yurovsky, D. (2018). A communicative approach to early word learning. *New Ideas in Psychology*, *50*, 73–79.

Yurovsky, D., Doyle, G., & Frank, M. C. (2016). Linguistic input is tuned to children's developmental level. In *Proceedings of the annual meeting of the cognitive science society* (pp. 2093–2098).

Yurovsky, D., Fricker, D. C., Yu, C., & Smith, L. B. (2014). The role of partial knowledge in statistical word learning. *Psychonomic Bulletin & Review, 21*, 1–22.

Yurovsky, D., Yu, C., & Smith, L. B. (2013). Competitive processes in cross-situational word learning. *Cognitive Science, 37*, 891–921.