A communicative framework for early word learning

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

[1] University of Chicago

[2] Carnegie Mellon University

Author Note

Correspondence concerning this article should be addressed to Benjamin C. Morris, Department of Psychology, University of Chicago, 5848 S University Ave, Chicago, IL 60637. E-mail: yurovsky@uchicago.edu

Abstract

Children do not learn language from passive observation of the world, but from interaction with caregivers who want to communicate with them. These communicative exchanges are structured at multiple levels in ways that support support language learning. We argue this pedagogically supportive structure can result from pressure to communicate successfully with a linguistically immature partner. We first characterize one kind of pedagogically supportive structure in a corpus analysis: caregivers provide more information-rich referential communication, using both gesture and speech to refer to a single object, when that object is rare and when their child is young. Then, in an iterated reference game experiment on Mechanical Turk (n = 480), we show how this behavior can arise from pressure to communicate successfully with a less knowledgeable partner. Lastly, we show that speaker behavior in our experiment can be explained by a rational planning model, without any explicit teaching goal. We suggest that caregivers' desire to communicate successfully may play a powerful role in structuring children's input in order to support language learning.

Keywords: language learning; communication; computational modeling

Word count: X

A communicative framework for early word learning

Introduction

One of the most striking aspects of children's language learning is just how quickly they master the complex system of their natural language (Bloom, 2000). In just a few short years, children go from complete ignorance to conversational fluency in a way that is the envy of second-language learners attempting the same feat later in life (Newport, 1990). What accounts for this remarkable transition?

Distributional learning presents a unifying account of early language learning: Infants come to language acquisition with a powerful ability to learn the latent structure of language from the statistical properties of speech in their ambient environment (Saffran, 2003). Distributional learning mechanisms can be seen in accounts across language including phonemic discrminitation (Maye, Werker, & Gerken, 2002), word segmentation (Saffran, 2003), learning the meanings of both nouns (Smith & Yu, 2008) and verbs (Scott & Fisher, 2012), learning the meanings of words at multiple semantic levels (Xu & Tenenbaum, 2007), and perhaps even the grammatical categories to which a word belongs (Mintz, 2003). A number of experiments clearly demonstrate both the early availability of distributional learning mechanisms and their potential utility across these diverse language phenomena (DeCasper & Fifer, 1980; DeCasper & Spence, 1986; Gomez & Gerken, 1999; Graf Estes, Evans, Alibali, & Saffran, 2007; Maye, Werker, & Gerken, 2002; Saffran, Newport, & Aslin, 1996; Smith & Yu, 2008; Xu & Tenenbaum, 2007).

However, there is reason to be suspicious about just how precocious statistical learning abilities are in early development. Although these abilities are available early, they are highly constrained by limits on other developing cognitive capacities. For example, infants' ability to track the co-occurrence information connecting words to their referents is constrained significantly by their developing memory and attention systems (Smith & Yu, 2013; Vlach & Johnson, 2013). Computational models of these processes show that the rate 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 listeners engage in recursive reasoning to produce and interpret speech cues by making inferences over one another's intentions (Frank & Goodman, 2012). These accounts have made room for advances in our understanding of a range of language phenomena previously uncaptured by formal modeling, notably a range of pragmatic inferences (e.g., Frank & Goodman, 2012; other RSA papers). In this work, we consider the communicative structure that emerges from an optimal communication system across a series of interactions where one partner has immature linguistic knowledge. This perspective offers the first steps toward a unifying account of both the child's learning and the parents' production: Both are driven by a pressure to communicate successfully (Brown, 1977).

Early, influential functionalist accounts of language learning focused on the importance of communicative goals (e.g., Brown, 1977). Our goal in this work is to formalize the intuitions in these accounts in a computational model, and to test this model against experimental data. We take as the caregiver's goal the desire to communicate with the child, not about language itself, but instead about the world in front of them. To succeed, the caregiver must produce the kinds of communicative signals that the child can understand and respond contingently, potentially leading caregivers to tune the complexity of their speech as a byproduct of in-the-moment pressure to communicate successfully (Yurovsky, 2017).

To examine this hypothesis, we focus on ostensive labeling (i.e. using both gesture and speech in the same referential expression) as a case-study phenomenon of information-rich structure in the language learning environment. We first analyze naturalistic parent communicative behavior in a longitudinal corpus of parent-child interaction in the home (Goldin-Meadow et al., 2014). We investigate the extent to which parents tune their ostensive labeling across their child's development to align to their child's developing linguistic knowledge (Yurovsky, Doyle, & Frank, 2016).

We then experimentally induce this form of structured language input in a simple 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– an information-rich cue that we take as one instance of pedagogically supportive language input. While many aspects of CDS support learning, multi-modal cues (e.g., speaking while pointing or looking) are uniquely powerful sources of data for young children (e.g., Baldwin, 2000). 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, Case, & Frank, 2017).

Methods

We used data from the Language Development Project– a large-scale, longitudinal corpus of parent child-interaction in the home with 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. Recordings were 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.

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 (in specific noun form), gesture (only deictic gestures were coded for ease of coding and interpretation– e.g., pointing) or both simultaneously.

Participants.

Results

These corpus data were analyzed using a mixed effects regression to predict parent use of multi-modal reference for a given referent. Random effects of subject and referent were included in the model. Our key predictors were child age and logged referent frequency (i.e. how often a given object was referred to overall across our data).

We find a significant negative effect of child age (in months) on multi-modal reference, such that parents are significantly less likely to produce the multi-modal cue as their child gets older (B = -0.04, $p < \mathit{0.0001}$). We also find a significant negative effect of referent frequency on multi-modal reference as well, such that parents are significantly less likely to provide the multi-modal cue for frequent referents than infrequent ones (B = -0.13, $p < \mathit{0.0001}$). 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.

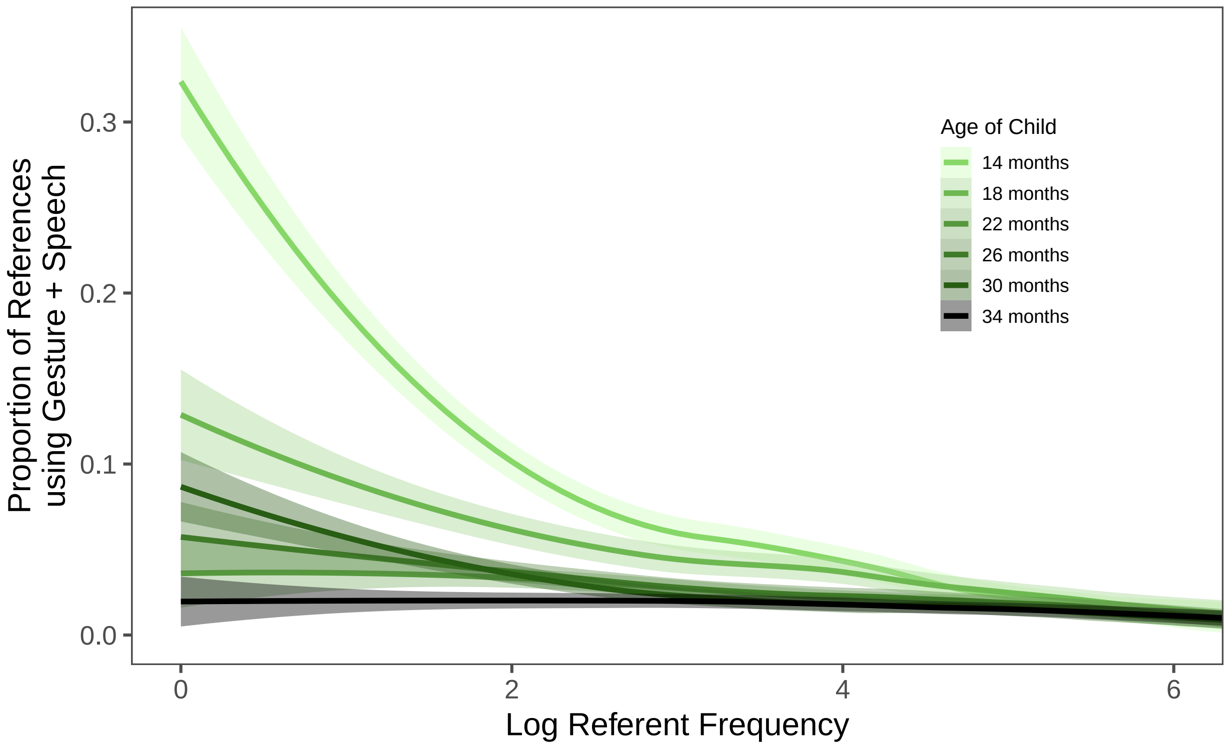

Figure 1. (#fig:corpus_plot)Proportion of parent multi-modal referential talk across development. The log of a referent's frequency is given on the x-axis, with less frequent items closer to zero.

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

We developed a simple reference game in which participants would be motivated to communicate successfully on a trial-by-trial basis. In all conditions, participants were placed in the role of speaker and asked to communicate with a computerized listener whose responses were programmed to be contingent on speaker behavior. We manipulated the relative costs of the communicative methods (gesture and speech) across conditions, as we did not have a direct way of assessing these costs in our naturalistic data, and they may 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 also established knowledge asymmetries by pre-training participants and manipulating how much training they thought their partner received. Using these manipulations, we aimed to experimentally determine the circumstances under which richly-structured input emerges, without an explicit pedagogical goal.

Experiment 1

Method

Participants. 480 participants were recruited though Amazon Mechanical Turk and received $1 for their participation. Data from 51 participants were excluded from subsequent analysis for failing the critical manipulation check and a further 28 for producing pseudo-English labels (e.g., "pricklyyone"). The analyses reported exclude the data from those participants, but all analyses were also conducted without excluding any participants and all patterns hold ($ps < 0.05$).

Design and Procedure. Participants were exposed to nine novel objects, each with a randomly assigned pseudo-word label. We manipulated the exposure rate within-subjects: during training participants saw three of the nine object-label mappings four times, two times, or one time. Participants were then given a recall task to establish their knowledge of the novel lexicon (pretest).

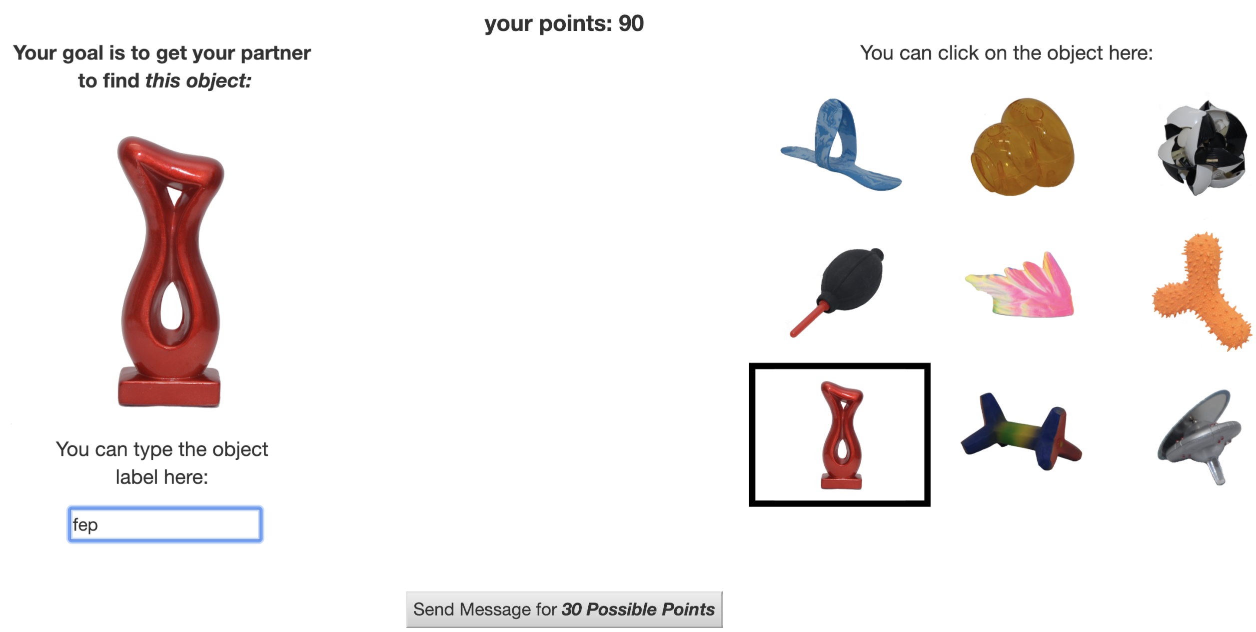

Figure 2. (#fig:exp_screenshot)Screenshot of speaker view during gameplay.

Prior to beginning the game, participants are told how much exposure their partner has had to the lexicon and also that they will be asked to discuss each object three times. As a manipulation check, participants are then asked to report their partner's level of exposure, and are corrected if they answer wrongly. Then during gameplay, speakers saw a target object in addition to an array of all nine objects (see Figure **??** for the speaker's perspective). Speakers had the option of either directly click on the target object in the array (gesture)- a higher cost cue but without ambiguity- or typing a label for the object (speech)- a lower cost cue but contingent on the listener's shared linguistic knowledge. After sending the message, speakers are shown which object the listener selected.

Speakers could win up to 100 points per trial if the listener correctly selected the target referent. We manipulated the relative utility of the speech cue between-subjects across two conditions: low relative cost for speech ("Low Relative Cost") and higher relative cost for speech (“Higher Relative Cost”). In the “Low Relative Cost” condition, speakers were charged 70 points for gesturing and 0 points for labeling, yielding 30 points and 100 points respectively if the listener selected the target object. In the “Higher Relative Cost” condition, speakers were charged 50 points for gesturing and 20 points for labeling, yielding up to 50 points and 80 points respectively. If the listener failed to identify the target object, the speaker nevertheless paid the relevant cost for that message in that condition. As a result of this manipulation, there was a higher relative expected utility for labeling in the “Low Relative Cost” condition than the “Higher Relative Cost” condition.

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).

To explore the role of listener knowledge, we also manipulated participants’ expectations about their partner’s knowledge across 3 conditions. Participants were told that their partner had either no experience with the lexicon, had the same experience as the speaker, or had twice the experience of the speaker.

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.

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

Results

As an initial check of our exposure manipulation, we fist a logistic regression predicting accuracy at test from a fixed effect of exposure rate and random intercepts and slopes of exposureRate by participant as well as random intercepts by item. We found a reliable effect of exposure rate, indicating that participants were better able to learn items that appear more frequently in training ($\beta = 1.09$, $t = 13.73$, $p < .001$). On average, participants knew at least 6 of the 9 words in the lexicon (mean = 6.28, sd = 2.26).

Gesture-Speech Tradeoff. To determine how gesture and speech are trading off across conditions, we looked at a mixed effects logistic regression to predict whether speakers chose to produce a label during a given trial as a function of the exposure rate, object instance in the game (first, second, or third), utility manipulation, and partner manipulation. A random subjects effects term was included in the model. There was a significant effect of exposure rate such that there was more labeling for objects with two exposures (B = NA, p

< 0.0001) or with four exposures (B = NA, $p < 0.0001$), compared with objects seen only once at training. Compared with the first instance of an object, speakers were significantly more likely to produce a label on the second appearance (B = 0.20, $p < 0.01$) or third instance of a given object (B = 0.46, $p < 0.0001$). Participants also modulated their communicative behavior on the basis of the utility manipulation and our partner exposure manipulation. Speakers in the Low Relative Cost condition produced significantly more labels than participants in the Higher Relative Cost condition (B = -0.84, $p < 0.001$). Speakers did more labeling with more knowledgeable partners; compared with the listener with no exposure, there were significantly higher rates of labeling in the same exposure (B = 1.74, $p < 0.0001$) and double exposure conditions (B = 3.13, $p < 0.001$).

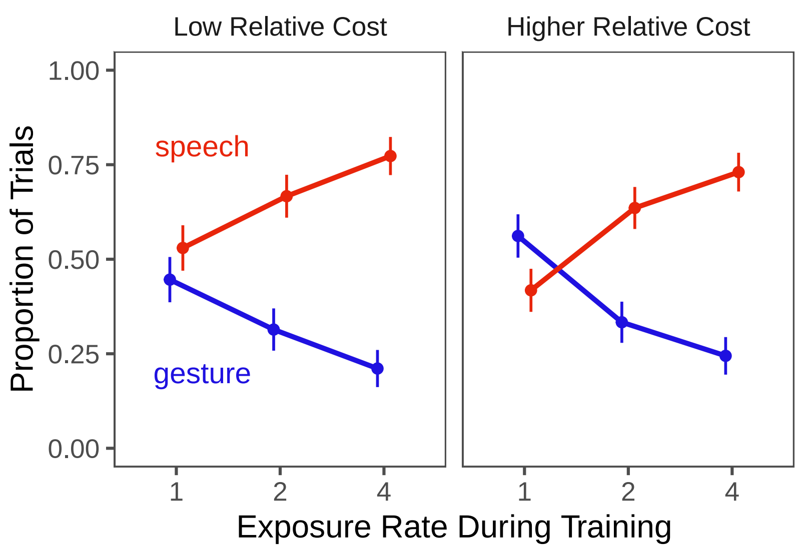

Figure 3. (#fig:speech_gesture)Speaker communicative method choice as a function of exposure and the utility manipulation.

Figure **??** illustrates the gesture-speech tradeoff pattern in the Double Exposure condition (as there was minimal teaching in that condition, so the speech-gesture trade-off is most interpretable). The effects on gesture mirror those found for labeling and are thus not included for brevity ($ps < 0.01$). Note that these effects cannot be explained by participant knowledge; all patterns above hold when looking *only* at words known by the speaker at pretest ($ps < 0.01$). Further, these patterns directly mirror 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.

In line with our hypotheses, a mixed effects logistic regression predicting whether or not teaching occurred on a given trial revealed that teaching rates across conditions depend on all of the same factors that predict speech and gesture (see Figure **??**). There was a significant positive effect of initial training on the rates of teaching, such that participants were more likely to teach words with two exposures (B = NA, $p < 0.05$) and four exposures (B = NA, $p < 0.05$), compared with words seen only once at training. There was also a significant effect of the utility manipulation such that being in the Low Relative Cost condition predicted higher rates of teaching than being in the Higher Relative Cost condition (B = -0.96, $p < 0.001$), a rational response considering teaching allows one to use a less costly strategy in the future and that strategy is especially superior in the Low Relative Cost condition.

We found an effect of partner exposure on rates of teaching as well: participants were significantly more likely to teach a partner with no prior exposure to the language than a partner with the same amount of exposure as the speaker (B = -1.63, $p < 0.0001$) or double their exposure (B = -3.51, $p < 0.0001$). The planned utility of teaching comes from using another, cheaper strategy (speech) on later trials, thus the expected utility of teaching should decrease when there are fewer subsequent trials for that object, predicting that teaching rates should drop dramatically across trials for a given object. Compared with the first trial for an object, speakers were significantly less likely to teach on the second trial (B

= -0.84, $p < 0.0001$) or third trial (B = -1.67, $p < 0.0001$).

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.

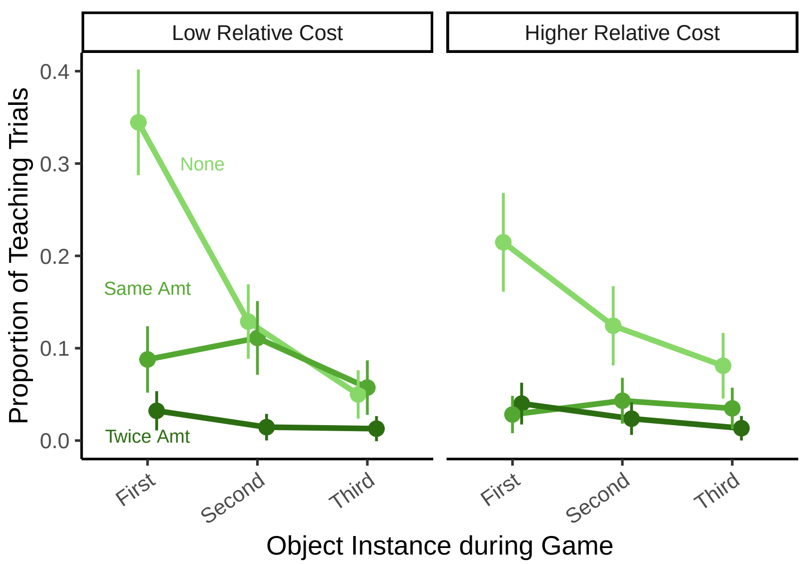

Figure 4. (#fig:exp_teach)Rates of teaching across the 6 conditions, plotted by how many times an object had been the target object.

Model: Communication as planning

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.

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 conversational partner. We can think of communication, then as a case of recursive planning. However, people do not have perfect knowledge of each-other's vocabularies (v). Instead, they only have uncertain beliefs (b) about these vocabularies that combine their expectations about what kinds of words people with as much linguistic experience as their partner are likely to know with their observations of their partner's behavior in past communicative interactions. This makes communication a kind of planning under uncertainty well modeled as a Partially Observable Markov Decision Process (POMDP, Kaelbling, Littman, & Cassandra, 1998).

Optimal planning in a POMDP involves a cycle of four phases: (1) Plan, (2) Act, (3) Observe, (4) Update beliefs. When people plan, they compute the Expected Utility of each possible action (a) by combining the Expected Utility of that action now with the Discounted Expected Utility they will get in all future actions. The amount of discounting (γ) reflects how people care about success now compared to success in the future. In our simulations, we set $\gamma = .5$ in line with prior work. Because Utilities depend on the communicative partner's vocabulary, people should integrate over all possible vocabularies in proportion to the probability that their belief assigns to that ($\mathbb{E}_{v \sim b}$).

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

Next, people take an action as a function of its Expected Utility. Following other models in the Rational Speech Act framework, we use the Luce Choice Axiom, in which each choice is taken in probability proportional to its exponentiated utility (Frank & Goodman, 2012; Luce, 1959). This choice rule has a single parameter α that controls the noise in this choice–as α approaches 0, choice is random and as α approaches infinity choice is optimal. For the results reported here, we set $\alpha = 2$ based on hand-tuning, but other values produce similar results.

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

After taking an action, people observe (o) their partner's choice–sometimes they pick the intended object, and sometimes they don't. 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.

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

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.

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

The final detail is to specify how people estimate their partner's learning rate (p) and initial vocabulary (v). We propose that people begin by estimating their own learning rate by reasoning about the words they learned at the start of the task: Their p is the rate that maximizes the probability of them having learned their initial vocabularies from the trials they observed. People can then expect their partner to have a similar p (per the "like me" hypothesis, Meltzoff, 2005). Having an estimate of their partner's p, they can estimate their vocabulary by simulating their learning from the amount of training we told them their partner had before the start of the game.

Model Results

The fit between our model's predictions and our empirical data from our reference game study on Amazon Turk can be seen in Figure **??**. The model outputs trial-level action predictions (e.g., "speak") for every speaker in our empirical data. These model outputs were aggregated across the same factors as the empirical data: modality, appearance, partner's exposure, and utility condition. We see a significant correlation of our model predictions and our empirical data ($r = 0.94$, $p<0.0001$). Our model provides a strong fit for these data, supporting our conclusion that richly-structured language input could emerge from in-the-moment pressure to communicate, without a goal to teach.

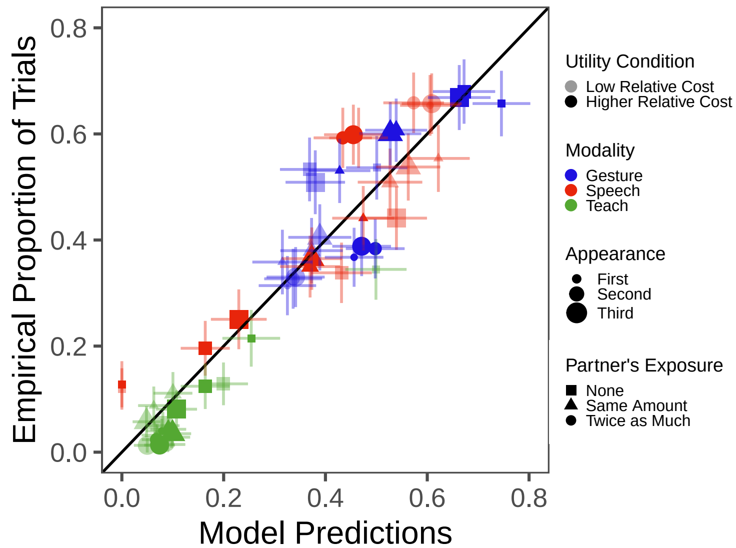

Figure 5. (#fig:model_fit)Fit between model predictions and empirical data.

Consequences for Learning

In the model and experiments above, we asked whether the pressure to communicate successfully with a linguistically-naive partner would lead to pedagogically supportive input. These results confirmed its' sufficiency: As long as linguistic communication is less costly than deictic gesture, speakers should be motivated to teach in order to reduce future communicative costs. Further, the strength of this motivation is modulated by predictable factors (speaker's linguistic knowledge, listener's linguistic knowledge, relative cost of speech and gesture, learning rate, etc.), and the strength of this modulation is well predicted by a 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 cost).

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; Yurovsky, Fricker, Yu, & Smith, 2014; although c.f. McMurray, 2007). Indeed, positive synergies across words are predicted by the majority of models and the impact of these synergies can be quite large under some assumptions about the frequency with which different words are encountered (Reisenauer, Smith, & Blythe, 2013). We assume independence primarily for pragmatic reasons here–it makes the simulations significantly more tractable (although it is what our experimental participants appear to assume about learners). Nonetheless, it is an important issue for future consideration. Of course, synergies that support learning under a cross-situational scheme must also support learning from communcators and teachers (Markman & Wachtel, 1988, @frank2009, @yurovsky2013). Thus, the ordering across conditions should remain unchanged. However, the magnitude of the difference sacross teacher conditions could potentially increase or decrease.

Method

Teaching. Because the teaching model is indifferent to communicative cost, it engages in ostensive an ostensive labeling (pointing + speaking) on each communicative event. Consequently, learning on each trial occurs with a probability that depends entirely on the learner's learning rate ($P_k = p$). Because we do not allow forgetting, the probability that a learner has failed to successfully learn after n trials is equal to the probability that they have failed to learn on each of n successive independent trials (The probabiliy of zero successess on n trials of a Binomial random variable with parameter p). The probability of learning after n trials is thus:

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

The expected probability of learning after n trials was thus defined analytically and required no simulation. For comparison to the other models, we computed P_k for values of p that ranged from .1 to 1 in increments of .1.

Communication. To test learner under the communication model, we implemented the same model described in the paper above. However, because our interest was in understanding the relationship between parameter values and learning outcomes rather than inferring the parameters that best describe people's behavior, we made a few simplifying assumptions to allow many runs of the model to complete in a more practical amount of time. First, in the full model above, speakers begin by inferring their own learning parameters (P_s) from their observations of their own learning, and subsequently use their maximum likelihood estimate as a standin for their listener's learning parameter (P_l). Because this estimate will converge to the true value in expectation, we omit these steps and simply stipulate that the speaker correctly estimates the listener's learning parameter.

Second, unless the speaker knows apriori how many times they will need to refer to a particular referent, the planning process is an infinite recursion. However, each future step in the plan is less impactful than the previous step (because of exponential discounting), this infinite process is in practice well approximated by a relatively small number of recursive steps. In our explorations we found that predictions made from models which planned over 3 future events were indistinguishable from models that planned over four or more, so we simulated 3 steps of recursion[1]. Finally, to increase the speed of the simulations we re-implemented them in the R programming language. All other aspects of the model were identical.

Hypothesis Testing. The literature on cross-situational learning is rich with a variety of models that could broadly be considered to be "hypothesis testers." In an eliminative hypothesis testing model, the learner begins with all possible mappings between words and objects and prunes potential mappings when they are inconsistent with the data according to some principe. A maximal version of this model relies on the principle that 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

[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-ratinoality in @griffiths2015). This future work is outside the scope of the current project.

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 prunes any that are not. If no current hypotheses are consistent, it adds up to h new hypotheses each with probability p. The model has converged when it has pruned all but the one correct hypothesis for the meaning of a word. This model is most similar to the Propose but Verify model proposed in Trueswell et al. (2013), with the exception that it allows for multiple hypotheses. Because of the data generating process, storing prior disconfirmed

[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 learnin 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.

hypotheses (as in Stevens et al., 2017), or incrementing hypotheses consistent with some but not all of the data (as in Yu & Smith, 2012) has no impact on learner and so we do not implement it here. We note also that, as described in Yu and Smith (2012), hypothesis testing models can mimic the behavior of associative learning models given the right parameter settings (Townsend, 1990).

In contrast to the Teaching and Communication simulations, the behavior of the Hypothesis Testing model depends on which particular non-target objects are present on each naming event. We thus began each simulation by generating a copus of 100 naming events, on each sampling the correct target as well as (C-1) competitors from a total set of M objects. We then simulated a hypothesis tester learning over this set of events as described above, and recorded the first trial on which the learner converged (having only the single correct hypothesized mapping between the target word and target object). We repeated this process 1000 times for each simulated combination of $M = (16, 32, 64, 128)$ total objects, $C = (1, 2, 4, 8)$ objects per trial, $h = (1, 2, 3, 4)$ concurrent hypotheses, as the learning rate p varied from .1 to 1 in increments of .1.

General Discussion

Across naturalistic corpus data, experimental data, and model predictions, we see evidence that pressure to communicate successfully with a linguistically immature partner could fundamentally structure parent production. In our experiment, we showed that people tune their communicative choices to varying cost and reward structures, and also critically to their partner's linguistic knowledge–providing richer cues when partners are unlikely to know the language and many more rounds remain. These data are consistent with the patterns shown in our corpus analysis of parent referential communication and demonstrate that such pedagogically supportive input could arise from a motivation to maximize communicative success while minimizing communicative cost– no additional motivation to teach is necessary. In simulation, we demonstrate that such structure could have profound implications for child language learning, simplifying the learning problem posed by most distributional accounts of language learning.

Accounts of language learning often aim to explain its striking speed in light of the sheer complexity of the language learning problem itself. Many such accounts argue that simple (associative) learning mechanisms alone seem insufficient to explain the rapid growth of language skills and appeal instead to additional explanatory factors, such as the so-called language acquisition device, working memory limitations, word learning biases, etc. (e.g., Chomsky, 1965; Goldowsky & Newport, 1993; Markman, 1990). While some have argued for the simplifying role of language distributions (e.g., McMurray, 2007), these accounts largely focus on learner-internal explanations. For example, Elman (1993) simulates language learning under two possible explanations to intractability of the language learning problem: one environmental, and one internal. He first demonstrates that learning is significantly improved if the language input data is given incrementally, rather than all-at-once (Elman, 1993). He then demonstrates that similar benefits can arise from learning under limited working memory, consistent with the "less-is-more" proposal (Elman, 1993; Goldowsky & Newport, 1993). Elman dismisses the first account arguing that ordered input is implausible, while shifts in cognitive maturation are well-documented in the learner (Elman, 1993); however, our account's emphasis on changing calibration to such learning mechanisms suggests the role of ordered or incremental input from the environment may be crucial.

This is consonant with work in other areas of development,

Recent research on the infant's visual learning environment has found surprising consistency and incrementality that could be a powerful tool for visual learning. Notably, research using head mounted cameras has demonstrated 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 motivated, contingent structure in the environment and the related learning abilities (Smith et al., 2018).

By analogy, the work we have presented here aims to draw a similar argument here for the language environment, which is also demonstrably beneficial for learning and shifting across development. In the case of language, the contingencies between learner and environment are even clearer than visual learning. Structure in the language environment that is continually suited to changing learning mechanisms must come in large part from caregivers themselves, and communicative, functional pressures make the caregiver speech highly dependent on the learner. 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.

NOT LANGUAGE BROADLY, BUT LANGUAGE FOR SPECIFIC WORDS ETC

Explaining parental modification is a necessary condition for building a complete theory of language, but parental modification need not be a necessary condition for language learning and is certainly not a sufficient condition. Our argument is that the rate and ultimate attainment of language learners will vary substantially as a function of parental modification, and that describing the cause of this variability is a necessary feature of models of language learning.

Generalizability and Limitations. Our account aims to put these processes into explicit dialogue and think about parent production and child learning in the same system.

While we have focused on ostensive labeling as a case-study phenomenon, our account should reasonably extend to the changing structure found in other aspects of child-directed speech– though see below for important limitations to this extension. Some such phenomena will be easily accounted for: aspects of language that shape communicative efficiency should shift in predictable patterns across development.

While these aspects of parent production can be captured by our proposed framework, incorporating them will likely require altering aspects of our account and decisions about which alterations are most appropriate. For example, the exaggerated pitch contours seen in infant-directed speech could be explained by this account if we expand the definition of communicative success to include a goal like maintaining attention. Alternatively, one could likely accomplish the same goal by altering the cost and utility structure to penalize loss of engagement. Thus, while this account should generalize to other modifications found in child-directed speech, such generalizations will likely require non-trivial alterations to the extant structure of the framework.

Of course, not all aspects of language should be calibrated to the child's language development– only those that support communication. Thus, our account also provides an initial framework for explaining aspects of communication that would not be modified in child-directed speech: namely, aspects of communication that minimally effect communicative efficiency. In other words, communication goals and learning goals are not always aligned. For example, children frequently overregularize past and plural forms and mastering the proper tense endings (i.e. the learning goal) might be aided by feedback from parent (citation on overregularization). However, adults rarely provide corrective feedback for these errors (citation for lack of correction), perhaps because incorrect grammatical forms are often sufficient to allow for successful communication (i.e. the communicative goal). The degree of alignment between communication and learning goals should predict the extent to which a linguistic phenomenon is modified in child-directed speech. Fully establishing the degree to which modification is expected for a given language phenomena will likely require working through a number of limitations in the generalizability of the framework as it stands.

Some aspects of parent production are likely entirely unrepresented in our framework, such as aspects of production driven by speaker-side constraints. Furthermore, our account is formulated primarily around concrete noun learning and future work must address its viability in other language learning problems. We chose to focus on ostensive labeling as a case-study phenomenon because it is an undeniably rich information source for young language learners, however ostensive labeling varies substantially across socio-economic status and cross-linguistically (citation for SES + lang ostensive labeling). This is to be expected to the extent that parent-child interaction is driven by different goals (or goals given different weights) across these populations– variability in goals could give rise to variability in the degree of modification. Nonetheless, the generalizability of our account across populations remains unknown. Indeed, child-directed speech itself varies cross-linguistically, both in its features (citation) and quantity (citation). There is some evidence that CDS predicts learning even in cultures where CDS is qualitatively different and less prevalent than in American samples (Schneidman). Future work is needed to establish the generalizability of our account beyond the western samples studied here.

We see this account as building on established, crucial statistical learning skills– language data from overheard speech or distributional information writ large are undoubtedly helpful for some learning problems (e.g., phoneme learning). There is likely large variability in the extent to which statistical learning skills drive the learning for a given learning problem. The current framework is limited by its inability to account for such differences across learning problems, which could derive from domain or cultural differences. Understanding generalizability of this sort and the limits of statistical learning will likely require a full account spanning both parent production and child learning.

A full account that explains variability in modification across aspects of language will rely on a fully specified model of optimal communication. Such a model will allow us to determine both which structures are predictably unmodified, and which structures must be modified for other reasons. Nonetheless, this work is an important first step in validating the hypothesis that language input that is structured to support language learning could arise from a single unifying goal: The desire to communicate effectively.

Conclusion

Acknowledgement

The authors are grateful to XX and YY for their thoughtful feedback on this manuscript. This research was supported by a James S MacDonnel Foundation Scholars Award to DY.

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.

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.

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.

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.

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.

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.

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.

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., Case, S., & Frank, M. C. (2017). Preschoolers flexibly adapt to linguistic input in a noisy channel. *Psychological Science*.

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.