From partners to populations: A hierarchical Bayesian account of coordination and convention

Robert D. Hawkins*[1], Michael Franke[2], Michael C. Frank[3],
Adele E. Goldberg[1], Kenny Smith[4], Thomas L. Griffiths[1,5], Noah D. Goodman[3,6]
[1]Department of Psychology, Princeton University, [2]Institute for Cognitive Science, University of Osnabrück,
[3]Department of Psychology, Stanford University, [4]Centre for Language Evolution, University of Edinburgh,
[5]Department of Computer Science, Princeton University, [6]Department of Computer Science, Stanford University

Languages are powerful solutions to coordination problems: they provide stable, shared expectations about how the words we say correspond to the beliefs and intentions in our heads. Yet language use in a variable and non-stationary social environment requires linguistic representations to be flexible: old words acquire new *ad hoc* or partner-specific meanings on the fly. In this paper, we introduce a hierarchical Bayesian theory of convention formation that aims to reconcile the long-standing tension between these two basic observations. More specifically, we argue that the central computational problem of communication is not simply transmission, as in classical formulations, but *learning* and *adaptation* over multiple timescales. Under our account, rapid learning within dyadic interactions allows for coordination on partner-specific common ground, while social conventions are stable priors that have been abstracted away from interactions with multiple partners. We present new empirical data alongside simulations showing how our model provides a cognitive foundation for explaining several phenomena that have posed a challenge for previous accounts: (1) the convergence to more efficient referring expressions across repeated interaction with the same partner, (2) the gradual transfer of partner-specific common ground to novel partners, and (3) the influence of communicative context on which conventions eventually form.

Keywords: Keywords TBD

To communicate successfully, speakers and listeners must share a common system of semantic meaning in the language they are using. These meanings are *social conventions* in the sense that they are arbitrary to some degree, but sustained by stable expectations that each person holds about others in their community (Lewis, 1969; Bicchieri, 2006; Hawkins, Goodman, & Goldstone, 2019). Importantly, these expectations extend to complete strangers. An English speaker may order a "cup of coffee" at any café in the United States and expect to receive (roughly) the same kind of drink.

At the same time, meaning can be remarkably flexible and *partner-specific*. The same words may be interpreted differently by different listeners. Interactions between friends and colleagues are filled with proper names, technical jargon, slang, shorthand, and inside jokes, many of which are unintelligible to outside observers. Furthermore, words take on new *ad hoc* senses over the course of a conversation (Clark, 1996).

The tension between these two basic observations, global stability and local flexibility, has posed a chal-

This report is based in part on work presented at the 39th, 40th, and 42nd Conferences of the Cognitive Science Society (Hawkins, Frank, & Goodman, 2017; Hawkins, Franke, Smith, & Goodman, 2018; Hawkins, Goodman, Goldberg, & Griffiths, 2020). Materials and code for reproducing all model simulations, behavioral experiments, and analyses are open and available online at https://github.com/hawkrobe/conventions_model.
*Correspondence should be addressed to Robert Hawkins, e-mail: rdhawkins@princeton.edu

lenging and persistent puzzle for theories of convention. Many influential computational accounts explaining how stable social conventions emerge in populations do not allow for partner-specific meaning at all (e.g. Hurford, 1989; Barr, 2004; Skyrms, 2010; Steels, 2011; Young, 2015). These accounts typically examine groups of interacting agents who update their representation of language after each interaction. While the specific update rules range from simple associative mechanisms (e.g. Steels, 1995) or heuristics (e.g Young, 1996) to more sophisticated deep reinforcement learning algorithms (e.g. Tieleman, Lazaridou, Mourad, Blundell, & Precup, 2019; Graesser, Cho, & Kiela, 2019; Mordatch & Abbeel, 2017), all of these accounts assume that agents update a single, monolithic representation of language to be used with every partner, and that agents do not (knowingly) interact repeatedly with the same partner.

Conversely, accounts emphasizing rapid alignment (Pickering & Garrod, 2004) or partner-specific common ground (Clark & Wilkes-Gibbs, 1986) across extended interactions with the same partner typically do not specify mechanisms by which community-wide conventions may arise over longer timescales. The philosopher Donald Davidson articulated one of the most radical of these accounts. According to Davidson (1984, 1986, 1994), while we bring in background expectations ("prior theories") it is the ability to coordinate on *partner-specific* meanings ("passing theories") that is ultimately responsible for communicative success:

> *In order to judge how he will be interpreted, [the speaker] uses a picture of the interpreter's readiness to interpret along certain lines, [...] the starting theory of interpretation. As speaker and interpreter talk, their "prior" theories become more alike; so do their "passing" theories. The asymptote of agreement and understanding is when passing theories coincide. Not only does it have its changing list of proper names and gerrymandered vocabulary, but it includes every successful use of any other word or phrase, no matter how far out of the ordinary [...] Such meanings, transient though they may be, are literal. (Davidson, 1986, p. 261).*

This line of argument led Davidson (1986) to memorably conclude that "there is no such thing as a language" (p. 265), and to abandon appeals to convention altogether (see also Heck, 2006; Lepore & Ludwig, 2007; Hacking, 1986; Dummett, 1994, for discussion).

In this paper, we propose a theory of coordination and convention that aims to reconcile the emergence of community-level conventions with partner-specific common ground in a unified cognitive model. This theory is motivated by the computational problems facing individual agents who must communicate with one another in a variable and non-stationary world. We suggest that three core cognitive capacities are needed for an agent to solve this problem, which are naturally formalized in a hierarchical Bayesian model:

C1: the ability to maintain **uncertainty** about what words will mean to different partners,

C2: flexible **online learning** to coordinate on partner-specific meanings, and

C3: inductive **generalization** to abstract away stable expectations about meaning to new partners.

One of our central theoretical aims is to ground the problem of convention — a fundamentally interactive, social phenomenon — in the same domain-general cognitive mechanisms supporting learning in other domains where abstract, shared properties need to be inferred along with idiosyncratic particulars of instances (Berniker & Kording, 2008; Goodman, Ullman, & Tenenbaum, 2011; Tenenbaum, Kemp, Griffiths, & Goodman, 2011; Kleinschmidt & Jaeger, 2015).

Our argument is structured around a series of three key phenomena in the empirical literature that have proved evasive for previous theoretical accounts of convention and coordination:

P1: the convergence to **increasingly efficient conventions** over repeated interactions with a single partner,

P2: the transition from **partner-specific conventions** to ones that are expected to generalize to new partners, and

P3: the influence of **communicative context** on which terms eventually become conventionalized

We begin by introducing the *repeated reference game* paradigm at the center of this literature and reviewing the empirical evidence supporting each of these phenomena. We then introduce our hierarchical Bayesian

model in detail and highlight several qualitative properties that emerge from our formulation. The remainder of the paper proceeds through each of these phenomena (**P1**-**P3**) in turn. For each phenomenon, we use computational simulations to evaluate our model's explanations on existing data, and introduce data from new real-time, multi-player behavioral experiments to test novel predictions when existing data does not suffice. Finally, we close by discussing several broader consequences of the theory, including the continuity of language acquisition in childhood with adaptation in adulthood, the domain-generality of discourse processes, and the integration of social and linguistic representations.

Three lessons about convention formation from repeated reference games

A core function of language is *reference*: using words to convey the identity of an entity in the environment. Loosely inspired by Wittgenstein (1953), empirical studies of coordination and convention in communication have predominantly focused on the subset of language use captured by simple "reference games." In a reference game task, participants are assigned to speaker and listener roles and shown a context of possible referential targets (e.g. images). On each trial, the speaker is asked to produce a referring expression — typically a noun phrase — that will allow the listener to select the intended target object from among the other objects in the context. Critically, a *repeated reference game* asks them to refer to the same targets multiple times as they build up a shared history of interaction with their partner (see Table A1 in Appendix for a review of different axes along which the design varies).

Unlike typical studies of referring expression generation (van Deemter, 2016; Degen, Hawkins, Graf, Kreiss, & Goodman, 2020; Dale & Reiter, 1995), studies of coordination and convention tend to use novel, ambiguous stimuli that participants do not already have strong conventions for. And unlike agent-based simulations of convention formation on large networks (e.g. Steels, 2011; Barr, 2004; Centola & Baronchelli, 2015), which typically match agents with a new, anonymous partner for each trial, repeated reference games ensure that participants maintain the same partner throughout an extended interaction. This design allows us to observe how the speaker's referring expressions for the same objects change as a function of interaction with that particular partner. We highlight three findings of particular theoretical significance in this literature.

P1: Increasingly efficient conventions. The most well-known phenomenon observed in repeated reference games is a dramatic reduction in message length over multiple rounds (Krauss & Weinheimer, 1964; Clark & Wilkes-Gibbs, 1986; Hawkins, Frank, & Goodman, 2020). The first time participants refer to a figure, they tend to use a lengthy, detailed description (e.g. "the upside-down martini glass in a wire stand") but with a small number of repetitions — between 3 and 6, depending on the pair of participants — the description may be cut down to the limit of just one or two words ("martini"). These final messages are as short or shorter than the messages participants produce when they are instructed to generate descriptions for themselves to interpret in the future (Fussell & Krauss, 1989) and are often incomprehensible to overhearers who were not present for the initial messages (Schober & Clark, 1989). These observations set up our first puzzle of *ad hoc* convention formation in dyads. How does a word or short phrase that would have been ineffective for communicating under the global conventions of a language take on local meaning over mere minutes of interaction?

P2: Partner-specific conventions. Because meaning is grounded in the evolving common ground shared with each partner, *ad hoc* conventions established over a history of interaction with one partner are not necessarily transferred to other partners (Metzing & Brennan, 2003; Weber & Camerer, 2003; Brown-Schmidt, 2009). For example, Wilkes-Gibbs and Clark (1992) paired participants for a standard repeated reference game, but after six rounds, one partner was asked to leave the room and replaced by a naive partner. Without partner-specific representations, we would expect speakers to continue using the short labels they had converged on with their first partner; instead, speakers reverted to the longer utterances they had initially used and coordinated on new *ad hoc* conventions with their new partner (see Fig. 1).

These effects raise our second puzzle: how do population-level conventions form in the presence of such strong partner-specificity? When are agents justified in transferring an *ad hoc* convention formed with one partner to a new, unseen partner? One important empirical clue was provided by Fay, Garrod, Roberts,

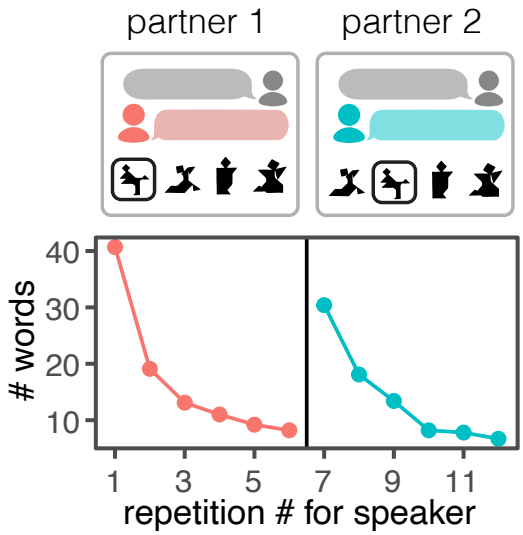

Figure 1. Classic phenomena in repeated reference games. Across multiple repetitions of the same referent with the same partner, speakers converge to increasingly efficient utterances (reps. 1-6). When the listener is replaced by a new, naive partner, speakers display a key signature of partner-specificity, reverting to longer utterances before converging again with their new partner (reps. 7-12). Error rates were reported to be low (~ 2.3%) throughout the experiment. Reproduced from Table 3 in Wilkes-Gibbs and Clark (1992).

and Swoboda (2010), who examined the emergence of conventions in a lab experiment where communities of eight people played a repeated graphical communication game similar to Pictionary, where participants produced drawings to allow their partner to identify a concept from a list of possibilities. The 8 participants in each network interacted dyadically with every other member of the community in, turn, for a series of seven repeated reference games.

Strikingly, participants behaved as observed by Wilkes-Gibbs and Clark (1992) at the first few partner swaps, consistent with partner-specificity, but the initial messages they sent to subsequent partners gradually converged closer to the *ad hoc* labels conventionalized with previous partners, indicating a slow gradient of generalization within their community. While intriguing, this work was limited by an extremely small sample size (N = 4 groups) and technical challenges facing the measurement of conventions in the graphical modality (see Hawkins, Sano, Goodman, & Fan, 2019). More recent work has adopted a similar design for an artificial-language communication task (Raviv, Meyer, & Lev-Ari, 2019) but has collapsed across repeated dyadic interactions to exclusively analyze network-level metrics, making it difficult to assess any effects of partner-specificity. Given these limitations of existing data, we evaluate our model's predictions using a large-scale, real-time web experiment directly extending Wilkes-Gibbs and Clark (1992) to larger networks.

P3: Context-sensitive conventions. Finally, while a degree of arbitrariness is central to conventionality – there must exist more than one solution that would work equally well – this does not necessarily imply that all possible conventions for a meaning are equally likely in practice, or even that all meanings are equally likely to become conventionalized in the first place (Hawkins & Goldstone, 2016). Indeed, functional accounts of language have frequently observed that lexical systems are well-calibrated to the needs of users under the statistics of their communicative environment (Gibson et al., 2019). This Optimal Semantic Expressivity hypothesis (OSE; Frank, 2017) has held remarkably well for the lexical distributions found in natural languages across semantic domains like color words and kinship categories (Kemp & Regier, 2012; Regier, Kemp, & Kay, 2015; Gibson et al., 2017; Kemp, Xu, & Regier, 2018).

While such long-term, diachronic sensitivity to context has been explained by abstract principles of optimality, such as the equilibria concepts of evolutionary game theory (Jäger, 2007; Jäger & Van Rooij, 2007), it has not yet been grounded in a cognitive and mechanistic account of the immediate, synchronic processes unfolding in the minds of individual agents while they interact. In other words, while there is abundant empirical evidence for context-sensitivity in the *outcomes* of convention formation processes, our third puzzle concerns which cognitive mechanisms may be necessary or sufficient to give rise to such conventions.

Repeated reference games have emerged as a promising method for probing these mechanisms in the lab. We can explicitly control the communicative context and observe the resulting distribution of conventions that emerge when participants communicate with artificial languages (Winters, Kirby, & Smith, 2014; Kirby, Tamariz, Cornish, & Smith, 2015; Winters, Kirby, & Smith, 2018) or natural language (Hawkins, Frank, & Goodman, 2020). While these outcomes are informative, it has remained challenging to directly evaluate cognitive models against the *full trajectories* of con-

vention formation on a trial-by-trial basis. In our final section, we report new empirical data from a dyadic repeated reference task manipulating context, where simulated agents and human participants are shown exactly the same sequence of trials.

Convention formation as Hierarchical Bayesian inference

In this section, we propose a unified computational account of *ad hoc* coordination and convention formation that addresses these three empirical puzzles. We begin from first principles: What is the core computational problem that must be solved to achieve successful communication? Classically, this problem has been formulated in terms of coding and compression (Shannon, 1948). An intended meaning in the speaker's mind must be encoded as a signal that is recoverable by the receiver after passing through a noisy transmission channel. This literal transmission problem has since been enriched to account for *pragmatics* – the ability of speakers and listeners to use context and social knowledge to draw inferences beyond the literal meaning of messages (Sperber & Wilson, 1986). We take the Rational Speech Act framework (RSA; Frank & Goodman, 2012; Goodman & Frank, 2016; Franke & Jäger, 2016) as representative of this current synthesis, formalizing communication as recursive social inference in a probabilistic model. In the next section, we will review this basic framework and then raise two fundamental computational problems facing this framework that motivate our proposal.

RSA models of communication with static meaning

In our referential communication setting[1], the RSA framework defines a pragmatic speaker, denoted by S_1, who must choose an utterance u that will allow their partner to choose a particular target object o from the current communicative context C: They attempt to satisfy the Gricean Maxims (Grice, 1975) by selecting utterances proportional to a utility function $U(u; o)$ that balances informativity to an imagined listener against the cost of producing an utterance:

$$S_1(u|o) \propto \exp\{\alpha_S \cdot U(u; o)\} \quad (1)$$
$$U(u; o) = (1 - w_C) \cdot \underbrace{\log L_0(o|u)}_{\text{informativity}} - w_C \cdot \underbrace{c(u)}_{\text{cost}}$$

where $c(u)$ is a function giving the cost of producing u, assuming a longer utterances are more costly. The speaker has two free parameters: $w_C \in [0, 1]$ controls the relative weight of informativity and parsimony in the speaker's production, and $\alpha_S \in [0, \infty]$ controls their softmax optimality (i.e. as $\alpha_S \to \infty$, the speaker increasingly chooses the utterance with maximal utility.)

The imagined *literal listener* L_0 in Eq. 1 is assumed to identify the target using a lexical meaning function $\mathcal{L}(u, o)$ capturing the literal semantics of the utterance u. That is, the probability of the literal listener choosing object o is proportional to the meaning of u under a static lexical meaning function $\mathcal{L}$:

$$L_0(o|u) \propto \mathcal{L}(u, o)$$

Throughout this paper, we will take $\mathcal{L}$ to be a traditional truth-conditional function evaluating whether a given object is in the extension of the utterance[2]:

$$\mathcal{L}(u, o) = \begin{cases} 1 & \text{if } o \in [\![u]\!] \\ 0 & \text{otherwise} \end{cases}$$

However, there are many alternative representational choices compatible with our core model, including fuzzier, continuous semantics (Degen et al., 2020) or vector embeddings learned by a neural network (Potts, 2019, see Appendix B for examples), which may be more appropriate for scaling the model to larger spaces of words and referents. We return to these possibilities in the General Discussion.

Two fundamental problems for static meaning

This basic framework and its extensions have accounted for a variety of important phenomena in pragmatic language use (e.g. Scontras, Tessler, & Franke, 2018; Kao, Wu, Bergen, & Goodman, 2014; Tessler & Goodman, 2018; Lassiter & Goodman, 2015). Yet it retains a key assumption from classical models: that the speaker and listener must share the same literal "protocol" $\mathcal{L}$ for encoding and decoding messages. In this

[1] For concreteness, we restrict our scope to reference in a discrete context of objects, but the same formulation applies to more general spaces of meanings.

[2] Note that the normalization constant may be exactly zero for some possible lexicons – for instance, if a given utterance is literally false of all objects in context – in which case these distributions are not well-defined. See Appendix A for technical details of how we address this problem.

section, we highlight two under-appreciated challenges of communication that complicate this assumption.

The first challenge is *variability* in linguistic meaning throughout a language community. Different listeners may recover systematically different meanings from the same message, and different speakers may encode the same message in different ways. For example, doctors may fluently communicate with one another about medical conditions using specialized terminology that is meaningless to a patient. The words may not be in the patient's lexicon, and even common words may be used in non-standard ways. That is, being fluent speakers of the same language does not ensure perfect overlap for the relevant meanings that need to be transmitted in every context: different partners may simply be using different functions $\mathcal{L}$.

The second challenge is the *non-stationarity* of the world. Agents are continually presented with new thoughts, feelings, and entities, which they may not already have efficient conventions to talk about. For example, when new technology is developed, the community of developers and early adopters must find ways of referring to the new concepts they are working on (e.g. *e-mailing*, *the Internet*). Or, when researchers design a new experiment with multiple conditions, they must find ways of talking about their own *ad hoc* abstractions, often converging on idiosyncratic names that can be used seamlessly in meetings. That is, any literal protocol $\mathcal{L}$ that we may write down at one time would be quickly outdated at a later time (see Lazaridou et al., 2021, for a demonstration of the related problems posed by non-stationary for large neural language models). We must have some ability to extend our language on the fly as needed.

A hierarchical model of dynamic meaning

Rather than assuming a monolithic, universally shared language, we argue that agents solve the core problems posed by variability and non-stationarity by attempting to continually, adaptively *infer* the system of meaning used by their partner in context. When all agents are continually learning in this way, we will show that they are not only able to locally coordinate on *ad hoc* meanings with specific partners but also able to abstract away linguistic conventions that are expected to be shared across an entire community. We introduce our model in three steps, corresponding to three core capacities: hierarchical uncertainty about meaning, online partner-specific learning, and inductive generalization.

C1: Hierarchical uncertainty about meaning. When an agent encounters a communication partner, they must call upon some representation about what they expect different signals will mean to that partner. We therefore replace the static function $\mathcal{L}$ with a *parameterized family* of lexical meaning functions by $\mathcal{L}_\phi$, where different values of ϕ yield different possible systems of meaning. To expose the dependence on a fixed system of meaning, Eq. 1 can be re-written to give behavior under a fixed value of ϕ:

$$
\begin{aligned}
L_0(o|u,\phi) &\propto \mathcal{L}_\phi(u,o) \\
U(u;o,\phi) &= (1-w_C)\cdot \log L_0(o|u,\phi) - w_C \cdot c(u) \\
S_1(u|o,\phi) &\propto \exp\{\alpha_S \cdot U(u;o,\phi)\}
\end{aligned}
\tag{2}
$$

While we will remain agnostic for now to the exact functional form of $\mathcal{L}_\phi$ and the exact parameter space of ϕ (see *Inference details* section below), there are two key computational desiderata we emphasize. First, given the challenge of variability raised in the previous section, these expectations ought to be sensitive to the overall statistics of the population. An agent should know that there is tighter consensus about the meaning of *dog* than the meaning of, say, specialized medical terms like *sclerotic aorta* (Clark, 1998), and conversely, should expect more consensus around how to refer to familiar concepts than new or ambiguous concepts. Second, this representation should also, in principle, be sensitive to the social identity of the partner: a cardiologist should have different expectations about a long-time colleague than a new patient.

The first desideratum, representing population variability, motivates a *probabilistic* formulation. Instead of holding a single static function $\mathcal{L}_\phi$, which an agent assumes is shared perfectly in common ground (i.e. one ϕ for the whole population), we assume each agent maintains *uncertainty* over the exact meaning of words as used by different partners. In a Bayesian framework, this uncertainty is specified by a prior probability distribution over possible values of ϕ. For example, imagine that under some possible values of ϕ, the term "sclerotic aorta" has truth conditions related to a specific condition of the heart, but under other values of ϕ, it does not: a well-trained doctor approaching a stranger should not assume their partner is using either ϕ but

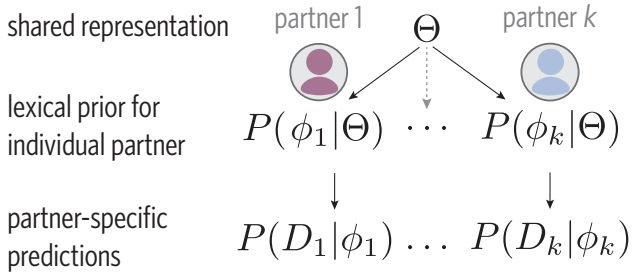

Figure 2. Schematic of hierarchical Bayesian model. At the highest level, denoted by Θ, is a representation of aspects of meanings expected to be shared across all partners. These *global* conventions serve as a prior for the systems of meanings used by specific partners, ϕ_k. These partner-specific representations give rise in turn to predictions about their language use $P(D_k|\phi_k)$, where D_k represents observations in a communicative interaction with partner k. By inverting this model, agents can adapt to *local* partner-specific conventions and update their beliefs about global conventions.

should assign some probability to each case. The introduction of uncertainty over a partner's literal semantics has previously been explored in the context of one-shot pragmatic reasoning, where it was termed *lexical uncertainty* (Bergen, Levy, & Goodman, 2016), and in the context of iterated dyadic interactions (Smith, Goodman, & Frank, 2013).

The second desideratum, sensitivity to partner-specific meanings, motivates a *hierarchical* model, where uncertainty is represented by a multi-level prior. At the highest level of the hierarchy is *community-level* uncertainty $P(\Theta)$, where Θ represents an abstract "over-hypothesis" about the overall distribution of possible partners. Θ then parameterizes the agent's *partner-specific* uncertainty $P(\phi_k|\Theta)$, where ϕ_k represents the specific system of meaning used by partner k (see Fig. 2). We focus for simplicity on this basic two-layer hierarchy, but the model can be straightforwardly extended to representing uncertainty at intermediate layers of social structure, including whether partners belong to distinct sub-communities (e.g. represented by discrete latent variables) or varying along latent dimensions (e.g. represented by a topic mixture). We return to these possible extensions in the General Discussion.

To integrate this lexical uncertainty into our speaker and listener models, we assume they each act in a way that is expected to be successful *on average*, under likely values of ϕ_k (Smith et al., 2013). In other words, they sample actions by marginalizing over their own beliefs $P_S(\phi_k)$ or $P_L(\phi_k)$ about different meanings their partner k may be using.

$$\begin{aligned} L(o|u) &\propto \exp\left\{\alpha_L \int P_L(\phi_k) \log S_1(u|o,\phi_k)\ d\phi_k\right\} \\ S(u|o) &\propto \exp\left\{\alpha_S \int P_S(\phi_k) U(u;o,\phi_k)\ d\phi_k\right\} \end{aligned} \tag{3}$$

where $\alpha_S, \alpha_L \in [0,\infty]$ control the speaker's and listener's soft-max optimality, respectively[3].

C2: Partner-specific online learning. The formulation in Eq. 3 derives how agents ought to act under uncertainty about the lexicon being used by their partner, $P(\phi_k)$. But how do beliefs about their partner change over time? Although an agent may begin with significant uncertainty about the system of meaning their partner is using in the current context, further interactions provide useful information for reducing that uncertainty and therefore improving the success of communication. In other words, *ad hoc* convention formation may be re-cast as an inference problem. Given observations D_k from interactions with partner k, an agent can update their beliefs about their partner's latent system of meaning following Bayes rule:

$$\begin{aligned} P(\phi_k, \Theta|D_k) &\propto P(D_k|\phi_k,\Theta)P(\phi_k,\Theta) \\ &= P(D_k|\phi_k)P(\phi_k|\Theta)P(\Theta) \end{aligned} \tag{4}$$

This joint inference decomposes the partner-specific learning problem into two terms, a prior term $P(\phi_k|\Theta)P(\Theta)$ and a likelihood term $P(D_k|\phi_k)$. The prior term captures the idea that, in the absence of strong evidence of partner-specific language use, the agent ought to regularize toward their background knowledge of conventions: the aspects of meaning that all partners are expected to share in common. The likelihood term represents predictions about how a partner would use language in context under different underlying systems

[3] We denote L and S without a subscript because they are the only speaker and listener models we use in simulations throughout the paper – the subscripted definitions are internal constructs used to define these models – but in the terminology of the RSA framework they represent L_1- and S_1-level pragmatic agents with lexical uncertainty. We found that higher levels of recursion were not strictly necessary to derive the phenomena of interest, but L_n and S_n-level lexical uncertainty models may be generalized by replacing S_1 in the listener equation, and L_0 in the speaker's utility definition, with standard RSA definitions of $n-1$-level agents (e.g. Zaslavsky, Hu, & Levy, 2020).

of meaning (as specified in the *Referential Feedback* section below).

Importantly, the posterior obtained in Eq. 4 allows agents to explicitly maintain *partner-specific expectations*, as used in Eq. 3, by marginalizing over community-level uncertainty:

$$P(\phi_k|D_k) = \int_\Theta P(\phi_k, \Theta|D_k)d\Theta \tag{5}$$

This posterior can be viewed as the "idiolect" that has been fine-tuned to account for partner-specific common ground from previous interactions. We will show that when agents learn about their partner in this way, and adjust their own production or comprehension accordingly (i.e. Eq. 3), they are able to coordinate on stable *ad hoc* conventions.

C3: Inductive generalization to new partners. The posterior in Eq. 4 also provides an inductive pathway for partner-specific data to inform beliefs about community-wide conventions. Agents update their beliefs about Θ, using data accumulated from different partners, by marginalizing over beliefs about specific partners:

$$P(\Theta|D) = \int_\phi P(\phi, \Theta|D)d\phi \tag{6}$$

where $D = \bigcup_{k=1}^{N} D_k$, $\phi = \phi_1 \times \cdots \times \phi_N$, and N is the number of partners previous encountered. Intuitively, when multiple partners are inferred to use similar systems of meaning, beliefs about Θ shift to represent this abstracted knowledge: it becomes more likely that a novel partner in one's community will share it as well. Note that this population-level posterior over Θ not only represents what the agent has learned about the central tendency of the group's conventions, but also the *spread* or variability, capturing the notion that some word meanings may be more widespread than others.

The updated Θ should be used to guide the prior expectations an agent brings into a subsequent interactions with strangers. This transfer is sometimes referred to as "sharing of strength" or "partial pooling" because pooled data is smoothly integrated with domain-specific knowledge. This property has been key to explaining how the human mind solves a range of other difficult inductive problems in the domains of concept learning (Kemp, Perfors, & Tenenbaum, 2007; Tenenbaum et al., 2011), causal learning (Kemp et al., 2007; Kemp, Goodman, & Tenenbaum, 2010), motor control (Berniker & Kording, 2008), and speech perception (Kleinschmidt & Jaeger, 2015). A key consequence of such transfer hierarchical models is the "blessing of abstraction," (Goodman et al., 2011) where it is possible under certain conditions for beliefs about the community's conventions *in general* to outpace beliefs about the idiosyncracies of individual partners (Gershman, 2017). We return to this property in the context of language acquisition in the General Discussion.

Further challenges for convention formation

The formulation in the previous section presents the core of our theory. Here, we highlight several additional features of our model, which address more specific challenges raised by prior work on communication and which we will encounter in the simulations reported in the remainder of the paper. Our organization of these details is motivated by Spike, Stadler, Kirby, and Smith (2017), which recently distilled three common problems that all accounts of convention must address: (1) the availability of referential feedback, (2) a form of information loss or forgetting, and (3) a systemic bias against ambiguity. Finally, we explain practical details of how we perform inference in this model.

Referential feedback. Learning and adaptation depends on the availability and quality of observations D_k throughout a communicative interaction. If the speaker has no way of assessing the listener's understanding, or if the listener has no way of comparing their interpretation against the speakers intentions, however indirectly, they can only continue to rely on their prior, with no ground for conventions to form (Krauss & Weinheimer, 1966; Hupet & Chantraine, 1992; Garrod, Fay, Lee, Oberlander, & MacLeod, 2007). So, what data D_k should each agent use to update their beliefs at a particular point in an interaction?

In principle, we expect that D_k reflects all relevant sources of information that may expose an agent's understanding or misunderstanding, including verbal and non-verbal backchannels (*mmhmm*, nodding), clarification questions, and actions taken in the world. In the more minimal setting of a reference game, we use the full feedback provided by the task, where the speaker's intended target and the listener's response are revealed at the end of each trial. Formally, this information can be written as a set of tuples $D_k = \{o^*, u', o'\}_{t=1}^{T}$, where

o^* denotes the speaker's intended target, u' denotes the utterance they produced, and o' denotes the listener's response, on each previous trial t.

Now, to specify the likelihoods in Eq. 4 for our referential setting, we assume each agent should infer their partner's lexicon ϕ_k by conditioning on their *partner's* previous choices. The listener on a given trial should use the probability that a speaker would produce u to refer to the target o^* under different ϕ_k, i.e. $P_L(\{o^*, u', o'\}_t \,|\, \phi_k) = S_1(u'_t \,|\, o^*_t, \phi_k)$, and the speaker should likewise use the probability that their partner would produce response o' after hearing utterance u, $P_S(\{o^*, u', o'\}_t \,|\, \phi_k) = L_0(o'_t \,|\, u'_t)$, This symmetry, where each agent is attempting to learn from the other's behavior, creates a clear coordination problem[4]. In the case of an error, where the agent in the listener role hears the utterance u' and chooses an object o' other than the intended target o^*, they will receive feedback about the intended target and subsequently condition on the fact that the speaker chose u' to convey that target. Meanwhile, the agent in the speaker role will subsequently condition on the likelihood that the listener chose the object o' upon hearing their utterance. In other words, each agent will subsequently condition on slightly different data leading to conflicting beliefs. Whether or not agents are able to resolve early misunderstandings through further interaction and eventually reach consensus depends on a number of factors.

Memory and forgetting. One important factor is imposed by the basis cognitive mechanisms of memory and forgetting. It is unrealistic to expect that memories of every example of every past interaction in the set of observations D is equally accessible to the agent. Furthermore, this may be to the agent's advantage. Without any mechanism for forgetting, early errors may prevent coordination much later in an interaction, as each agent's lexical beliefs must explain all previous observations equally. Without the ability to discount earlier data points, agents may be prevented from ever reaching consensus with their partner (Spike et al., 2017).

Forgetting is typically incorporated into Bayesian models with a decay term in the likelihood function (Anderson & Schooler, 2000; Angela & Cohen, 2009; Fudenberg & Levine, 2014; Kalm & Norris, 2018).

$$P(D_k|\phi_k) = \prod_{\tau=0}^{T} \beta^\tau P(\{o^*, u', o'\}_{T-\tau} \,|\, \phi_k)$$

where $\tau = 0$ indexes the most recent trial T and decay increases further back through time. This decay term is motivated by the empirical power function of forgetting (Wixted & Ebbesen, 1991), and can be interpreted as the expectation over a process where observations have some probability of dropping out of memory at each time step. Indeed, this likelihood can be derived by simply extending our hierarchical model down an additional layer *within* each partner to allow for the possibility that they are using slightly different lexicons at different points in time; assuming a degree of autocorrelation between neighboring time points yields this form of discounting. Alternatively, at the algorithmic level, decay can be viewed as a form of weighted importance sampling, where more recent observations are preferentially sampled (Pearl, Goldwater, & Steyvers, 2010)[5]

Bias against ambiguity. A third specific challenge is posed by ambiguity: if a speaker uses a label to refer to one target, it is consistent with the data for the listener to subsequently believe that the same expression may be acceptable for other targets as well. In our account, this problem is naturally solved by the principles of *pragmatic reasoning* instantiated in the RSA framework (Grice, 1975), which has been explicitly linked to *mutual exclusivity* in word learning (Bloom, 2002; Frank, Goodman, & Tenenbaum, 2009; Smith et al., 2013; Gulordava, Brochhagen, & Boleda, 2020; Ohmer, König, & Franke, 2020). Gricean pragmatic reasoning critically allows agents to learn from the *absence* of evidence by reasoning about alternatives.

Pragmatic reasoning plays two distinct roles in our model. First, Gricean agents assume that their partner is using language in a cooperative manner and account for this when inferring their partner's language model. That is, we use these equations as the linking function

[4] In some settings, agents in one role may be expected to take on more of the burden of adaptation, leading to an asymmetric division of labor (e.g. Moreno & Baggio, 2014). This may be especially relevant in the presence of asymmetries in power, status, or capability. In principle, this could be reflected in differing values of parameters α_S and α_L, but we leave consideration of such asymmetries for future work.

[5] While this simple decay model is sufficient for our simulations, it is missing important mechanistic distinctions between working memory and long-term memory; for example, explaining convention formation over longer timescales may require an explicit model of consolidation.

in the likelihood $P(D_k|\phi_k)$, representing an agent's prediction about how a partner with meaning function ϕ_k would actually behave in context (Eq. 4). S_1 is used to learn from observations generated in the speaker role and L_1 is used to learn from observations generated in the listener role. For example, upon hearing their partner use a particular utterance u to refer to an object o, a pragmatic agent can not only infer that u means o in their partner's lexicon, but also that all other utterances u' likely do *not* mean o: if they did, the speaker would have used them instead. Second, agents do not only make passive inferences from observation, they participate in the interaction by *using language* themselves. A Gricean agent's own production and comprehension is also guided by cooperative principles (Eq. 3).

Many minor variations on the basic RSA model have been explored in previous work, and it is worth highlighting three technical choices in our formulation. First, both agents are "action-oriented," in the sense that they behave proportional to the utility of different actions, according to a soft-max normalization $\sigma(U(z)) = e^{U(z)}/\sum e^{U(z)}$. This contrasts with some RSA applications, where the listener is instead assumed to be "belief-oriented," simply inferring the speaker's intended meaning without producing any action of their own (Qing & Franke, 2015). Second, our instantiation of lexical uncertainty differs subtly from the one used by Bergen et al. (2016), which placed the integral over lexical uncertainty at a *single* level of recursion (specifically, within a pragmatic listener agent). Instead, we argue that it is more natural in an interactive, multi-agent setting for each agent to maintain uncertainty at the highest level, such that each agent is reasoning about their *partner*'s lexicon regardless of what role they are currently playing.

Lastly, we must define how the L_0 agent behaves when an utterance is false of all objects in context under their lexicon (i.e. when the normalization constant is exactly 0). Previous proposals have assumed that the L_0 should choose uniformly in this case. However, this assumption has the unintuitive consequence that S_1's utility of using an utterance known to be false of the target may be the same as an utterance known to be true. To address this concern, we always add a low-prior-probability 'null object' to the L_0's context, for which all utterances evaluate to true. This alternative can be interpreted as a 'failure to refer' and effectively prevents L_0 from assigning belief to a referent for which the utterance is literally false (this case is distinct from the case of a *contradiction*, which arises when defining the meaning of multi-word utterances in Section **P1** below.) See Appendix A for further details and discussion of our RSA implementation.

Inference details. While our simulations in the remainder of the paper each address different scenarios, we have aimed to hold as many details as possible constant throughout the paper. First, we must be concrete about the space of possible lexicons that parameterizes the lexical meaning function, $\mathcal{L}_\phi$. For consistency with previous Bayesian models of word learning (e.g. Xu & Tenenbaum, 2007) we take the space of possible meanings for an utterance to be the set of nodes in a concept taxonomy. When targets of reference are conceptually distinct, as typically assumed in signaling games, the target space of utterance meanings reduces to the discrete space of individual objects, i.e. $[\![u]\!]_\phi = \phi(u) \in O$ for all $u \in \mathcal{U}$. For this special case, the parameter space contains exactly $|O| \times |\mathcal{U}|$ possible values for ϕ, corresponding to all possible mappings between utterances and individual objects. Each possible lexicon can therefore be written as a binary matrix where the rows correspond to utterances, and each row contains one object. The truth-conditional function $\mathcal{L}_\phi(u, o)$ then simply checks whether the element in row u matches object o. For example, there are four possible lexicons for two utterances and two objects:

$$\phi \in \left\{ \begin{bmatrix} \blacksquare \\ \bullet \end{bmatrix}, \begin{bmatrix} \bullet \\ \blacksquare \end{bmatrix}, \begin{bmatrix} \bullet \\ \bullet \end{bmatrix}, \begin{bmatrix} \blacksquare \\ \blacksquare \end{bmatrix} \right\}$$

Second, having defined the support of the parameter ϕ, we can then define a simplicity prior $P(\phi) \propto \exp\{-|\phi|\}$ following Frank et al. (2009), where $|\phi|$ is the total size of each word's extension, summed across words in the vocabulary. Again, for traditional signaling games, this reduces to a uniform prior because all possible lexicons are the same size: $\phi(u_i) \sim \text{Unif}(O)$. We can compactly write distributions over ϕ in terms of the same utterance-object matrix, where row i represents the marginal distribution over possible meanings of utterance u_i. For example, the uninformative prior for two utterances and two objects can be written:

$$P(\phi) = \begin{bmatrix} \text{Unif}\{\blacksquare, \bullet\} \\ \text{Unif}\{\blacksquare, \bullet\} \end{bmatrix} = \begin{matrix} \blacksquare & \bullet \\ \begin{bmatrix} .5 & .5 \\ .5 & .5 \end{bmatrix} \end{matrix}$$

This prior becomes more important for **P3**, however, where we consider spaces of referents with more complex conceptual structure and a larger space of possible meanings. A single word may apply to multiple conceptually related referents or, conversely, may have an empty meaning, where it is effectively removed from the agent's vocabulary. In this more general case, a simplicity prior generically favors smaller word meanings as well as a smaller effective vocabulary size, since the null meaning has the smallest extension.

Finally, while the probabilistic model we have formulated in this section is theoretically well-motivated and mathematically well-defined, it has previously been challenging to actually derive predictions from it. Historically, interactive models have been challenging to study with closed-form analytical techniques and computationally expensive to study through simulation, likely contributing to the prevalence of simplified heuristics in prior work. Our work has been facilitated by recent advances in probabilistic inference techniques that have helped to overcome these obstacles. We have implemented our simulations in the probabilistic programming language WebPPL (Goodman & Stuhlmüller, electronic). All of our simulations iterate the following trial-level loop: (1) sample an utterance from the speaker's distribution, given the target object, (2) sample an object from the listener's object distribution, given the utterance produced by the speaker, (3) append the results to the list of observations, and (4) update both agents' posteriors, conditioning on these observations before continuing to the next trial.

To obtain the speaker and listener distributions (steps 1-2; Eq. 2), we always use exhaustive enumeration for exact inference. We would prefer to use enumeration to obtain posteriors over lexical meanings as well (step 4; Eq. 4), but as the space of possible lexicons ϕ grows, enumeration becomes intractable. For simulations related to **P2** and **P3**, we therefore switch to Markov Chain Monte Carlo (MCMC) methods to obtain samples from each agent's posteriors, and approximate the expectations in Eq. 3 by summing over these samples. Because we are emphasizing a set of phenomena where our model makes qualitatively different predictions than previous models, our goal in this paper is to illustrate and evaluate these *qualitative* predictions rather than provide exact quantitative fits to empirical data. As such, we proceed by examining predictions for a regime of parameter values (w_S, w_L, w_C, β) that help distinguish our predictions from other accounts, and leave the more computationally expensive task of empirical parameter estimation to future studies.

Phenomenon #1: *Ad hoc* conventions become more efficient

We begin by considering the phenomenon of increasing efficiency in repeated reference games: speakers use detailed descriptions at the outset but converge to an increasingly compressed shorthand while remaining understandable to their partner. While this phenomenon has been extensively documented, to the point of serving as a proxy for measuring common ground, it has continued to pose a challenge for models of communication.

For example, one possibility is that speakers coordinate on meaning through priming mechanisms at lower levels of representation, as proposed by influential *interactive alignment* accounts (Pickering & Garrod, 2004, 2006; Garrod & Pickering, 2009). While low-level priming may be at play in repeated reference tasks, especially when listeners engage in extensive dialogue or alternate roles, it is not clear why priming would cause descriptions to get shorter as opposed to aligning on the same initial description. Furthermore, priming alone cannot explain why speakers still converge to more efficient labels even when the listener is prevented from saying anything at all and only minimal feedback is provided showing that the listener is responding correctly (Krauss & Weinheimer, 1966); conversely, speakers continue using longer descriptions when they receive non-verbal feedback that the listener is repeatedly making errors (see also Hawkins, Frank, & Goodman, 2020). In these cases, there are no linguistic features available for priming or alignment mechanisms. Explaining when and why speakers believe that shorter descriptions will suffice requires a mechanism for coordination on meaning even given sparse, non-verbal feedback.

Another possibility is that speakers coordinate on meaning using some some lexical update rule that makes utterances more likely to be produced after communicative successes and less likely after communicative failures, such as a variant on the Roth-Erev reinforcement learning rule (Erev & Roth, 1998) adopted by a variety of agent-based models (Steels, 1995; Barr,

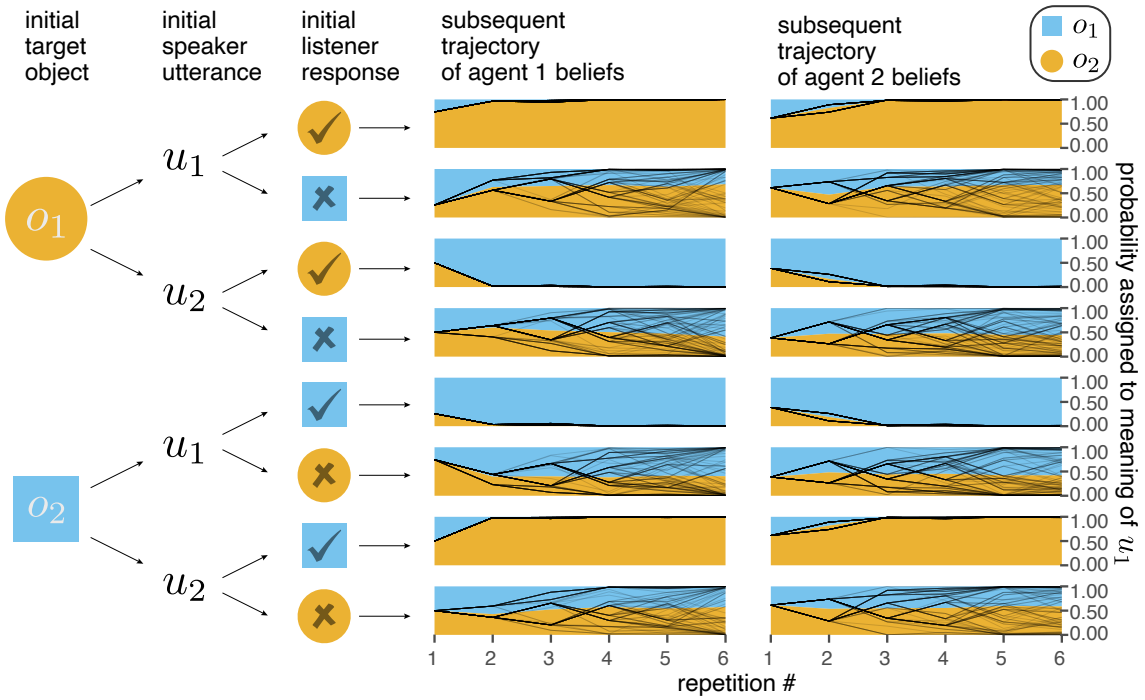

Figure 3. Path-dependence of conventions. The average trajectory of each agent's beliefs about the meaning of u_1, $\phi(u_1)$, is shown in blue and orange following all eight possible outcomes of the first trial in Simulation 1.1. For each of the two possible targets, the speaker could choose to produce either of the two utterances, and the listener could respond by choosing either of the two objects. In the cases where the listener chose correctly (marked with a checkmark), agents subsequently conditioned on the same data and rapidly converged on a system of meaning consistent with this feedback. For example, in the first row, when u_1 was successfully used to refer to the circle, both agents subsequently believe that u_1 means *circle* in their partner's lexicon. In the cases where the listener fails to choose the target, the agents subsequently condition on different data, and they converge on a convention that is determined by later choices (lines represent the trajectories of individual agents.)

2004; Young, 2015). While reinforcement is a powerful mechanism for allowing groups to reach consensus, it is not clear why a newly initialized speaker would prefer to produce longer utterances over shorter utterances, or, if this bias was built-in, how simply reinforcing initially long descriptions could lead utterances to get shorter. In the rare cases that some form of reduction has been investigated in this family of models (e.g. as in the phenomenon of phonological erosion), the process has been hard-coded as an ϵ probability of speakers dropping a random token at each point in time (Beuls & Steels, 2013; Steels, 2016).

Such random dropping, however, is an unsatisfying explanations for several reasons. First, it formalizes a reductive explanation of efficiency in terms of speaker-internal noise (or laziness). This idea dates back to the early literature on repeated reference games and control experiments by Hupet and Chantraine (1992) were designed to test this possibility (see also Garrod et al., 2007). Participants were asked to repeatedly refer to the same targets for a *hypothetical* partner to see later, such that any effects of familiarity or repetition on the part of the speaker were held constant with the interactive task. No evidence of reduction was found, and in some cases utterances actually grew longer. This accords with observations in multi-partner settings by Wilkes-Gibbs and Clark (1992), which we explore further in **P2**: it is difficult to explain why a speaker who only shortened their descriptions due to an ϵ-noise process would suddenly switch back to a longer utterance when their partner is exchanged. Whatever drives efficiency cannot be explained through speaker laziness, it must be a result of the *interaction* between partners.

In this section, we argue that our Bayesian account provides a rational explanation for increasing efficiency in terms of the inferences made by speakers across re-

peated interaction. Given that this phenomenon arises in purely dyadic settings, it also provides an opportunity to explore more basic properties of the first two capacities formalized in our model (representing *uncertainty* and *partner-specific learning*) before introducing hierarchical generalization in the next section. In brief, we show that increasing efficiency is a natural consequence of the speaker's tradeoff between informativity and parsimony (Eq. 3), given their inferences about the listener's language model. For novel, ambiguous objects like tangrams, where speakers do not expect strong referential conventions to be shared, longer initial descriptions are motivated by high initial uncertainty in the speaker's lexical prior $P(\phi_k|\Theta)$. Proposing multiple descriptors is a rational hedge against the possibility that a particular utterance will be misinterpreted and give the listener a false belief. As the interaction goes on, the speaker obtains feedback D_k from the listener responses and updates their posterior beliefs $P(\phi_k|D_k)$ accordingly. As uncertainty gradually decreases, they are able to achieve the same expected informativity with shorter, more efficient messages.

Simulation 1.1: Pure coordination

We build up to our explanation of increasing efficiency by first exploring a traditional signaling game scenario with only one-word utterances. This simulation tests the most fundamental competency for any model of *ad hoc* coordination: agents are able to coordinate on a communication system in the absence of shared priors. We consider the simplest possible reference game with two objects, $O = \{●, ■\}$, where the speaker must choose between two one-word utterances $\mathcal{U} = \{u_1, u_2\}$ with equal production cost.

We walk explicitly through the first step of the simulation to illustrate the model's dynamics (see Fig. 3). Suppose the target object presented to the speaker agent on the initial trial is ●. Both utterances are equally likely to apply to either object under the uniform lexical prior, hence each utterance is expected to be equally (un)informative. The speaker's utility therefore reduces to sampling an utterance at random $u \sim S(u|●)$. Suppose u_1 is sampled. The listener then hears this utterance and selects an object according to their own expected utility under their uniform lexical prior, which also reduces to sampling an object at random $o \sim L(o|u_1)$. Suppose they choose, ●, a correct response.

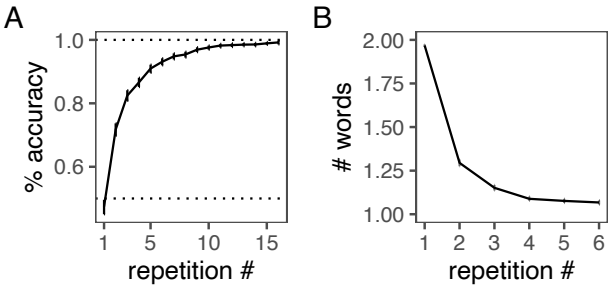

Figure 4. Pairs of agents learn to successfully coordinate on efficient ad hoc conventions over repeated interactions. (A) agents converge on accurate communication systems in Simulation 1.1, where only single-word utterances are available, and (B) converge on shorter, more efficient conventions in Simulation 1.2, where multi-word utterances were available. Error bars are bootstrapped 95% CIs across 1000 trajectories, computed within each repetition block of two trials.

Both agents may use the resulting tuple $D = \{●^*, u_1, ●\}$, depicted in the top row in Fig. 3 to update their beliefs about the lexicon their partner is using.

$$P_S(\phi|D) \propto L_0(● \,|\, u_1, \phi)P(\phi)$$
$$P_L(\phi|D) \propto S_1(u_1 \,|\, ●^*, \phi)P(\phi)$$

They then proceed to the next trial, where they use this updated posterior distribution to produce or interpret language instead of their prior. To examine how the dynamics of this updating process unfold over further rounds, we simulated 1000 such trajectories. The trial sequence was structured as a repeated reference game, containing 30 trials structured into 15 repetition blocks. The two objects appeared in a random order within each block, and agents swapped roles at the beginning of each block. We show representative behavior at softmax optimality parameter values $\alpha_L = \alpha_S = 8$ and memory discounting parameter $\beta = 0.8$, but find similar behavior in a wide regime of parameter values (see Appendix Fig. A1).

We highlight several key results from this simulation. First, and most fundamentally, the communicative success of the dyad rises over the course of interaction: the listener is able to more accurately select the intended target object (see Fig. 4A). Second, the initial symmetry between meanings in the prior is broken by initial choices, leading to *arbitrary* but *stable* mappings in future rounds. Because agents were initialized with the same priors in every trajectory, trajectories only diverged when different actions happen to

be sampled. This can be seen by examining the path-dependence of subsequent beliefs based on the outcome of the initial trial in Fig. 3. Third, we observe the influence of mutual exclusivity via Gricean pragmatic reasoning: agents also make inferences about objects and utterances that were *not* chosen. For example, observing $D = \{(●^*, u_2, ●)\}$ provides evidence that u_1 likely does not mean ● (e.g. the third row of Fig. 3, where hearing u_2 refer to ● immediately led to the inference that u_1 likely refers to ■).

Simulation 1.2: Increasing efficiency

Next, we show how our model explains speakers' gains in efficiency over multiple interactions. For efficiency to change at all, speakers must be able to produce utterances that vary in length. For this simulation, we therefore extend the model to allow for multi-word utterances by allowing speakers to combine together multiple primitive utterances. Intuitively, human speakers form longer initial description by combining a collection of simpler descriptions (e.g. "kind of an X, or maybe a Y with Z on top"). This raises a problem about how the meaning of a multi-word utterance $\mathcal{L}_\phi(u_1u_2)$ is derived from its components $\mathcal{L}_\phi(u_1)$ and $\mathcal{L}_\phi(u_2)$. To capture the basic desideratum that an object should be more likely to be chosen by L_0 when more components of the longer utterance apply to it, we adopt a standard conjunctive semantics[6]:

$$\mathcal{L}_\phi(u_iu_j, o) = \mathcal{L}_\phi(u_i, o) \times \mathcal{L}_\phi(u_j, o)$$

Now, we consider a scenario with the same two objects as in Simulation 1.1, but give the speaker four primitive utterances $\{u_1, u_2, u_3, u_4\}$ instead of only two, and allow two-word utterances such as u_1u_2. We established in the previous section that successful *ad hoc* conventions can emerge even in a state of pure uncertainty, but human participants in repeated reference games typically bring some prior expectations about language into the interaction. For example, a participant who hears 'ice skater' on the first round of the task in Clark and Wilkes-Gibbs (1986) may be more likely to select some objects more than others while still having substantial uncertainty about the intended target (e.g. over three of the twelve tangram that have some resemblance to an ice skater). We thus initialize both agents with weak biases δ (represented in compressed matrix form in Fig. 5):

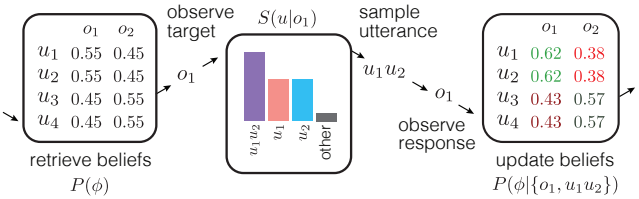

Figure 5. Schematic of speaker for first trial of Simulation 1.2. The speaker begins with uncertainty about the meanings in the listener's lexicon (e.g. assigning 55% probability to the possibility that utterance u_1 means object o_1.) A target o_1 is presented, and the speaker samples an utterance from the distribution $S(u|o_1)$. Finally, they observe the listener's response and update their beliefs. Due to the compositional semantics of the utterance u_1u_2, the speaker becomes increasingly confident that both component primitives, u_1 and u_2, apply to object o_1 in their partner's lexicon.

$$\phi(u_1), \phi(u_2) \sim \text{Categorical}(0.5 + \delta)$$
$$\phi(u_3), \phi(u_4) \sim \text{Categorical}(0.5 - \delta)$$

As in Simulation 1.1, we simulated 1000 distinct trajectories of dyadic interaction between agents. Utterance cost was defined to be the number of 'words' in an utterance, so $c(u_1) = 1$ and $c(u_1u_2) = 2$. As shown in Fig. 4B, our speaker agent initially prefers longer utterance (mean length $\approx$ 1.5 on first block) but rapidly converges to shorter utterances after several repetitions (mean length $\approx$ 1 on final block), qualitatively matching the curves measured in the empirical literature.

To illustrate in detail how our model derives this behavior, we walk step-by-step through a single trial (Fig.

[6] One subtle consequence of a conjunctive Boolean semantics is the possibility of contradictions. For example, under a possible lexicon where $\phi(u_1) =$ ■ and $\phi(u_2) =$ ●, the multi-word utterance u_1u_2 is not only false of a particular referent in the current context, it is false of all possible referents; it *fails to refer*, reflecting a so-called truth-gap (Strawson, 1950; Van Fraassen, 1966). We assume such an utterance is uninterpretable and simply does not change the literal listener L_0's beliefs. While this assumption is sufficient for our simulations, we regard this additional complexity as a limitation of classical Boolean semantics and show in Appendix C that switching to a more realistic continuous semantics with lexical values in $[0, 1]$ (Degen et al., 2020) may better capture the notion of redundancy that motivates speakers to initially produce longer utterances.

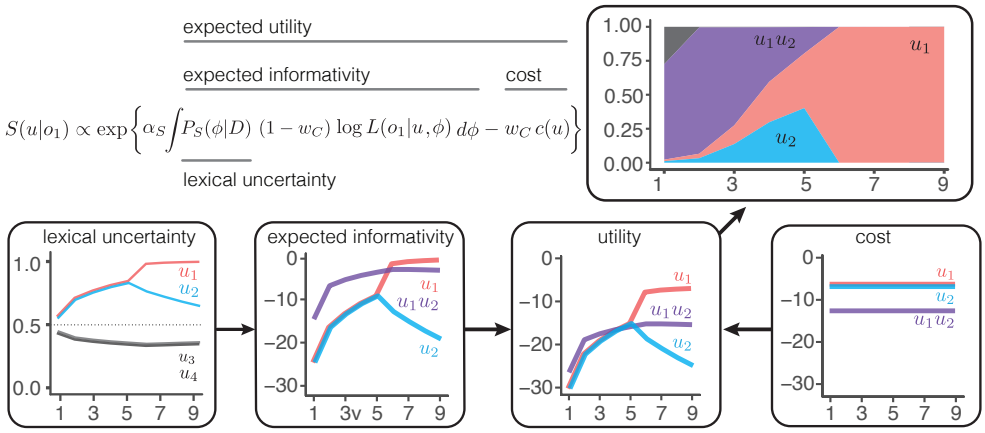

Figure 6. Internal state of speaker in example trajectory from Simulation 1.2. Each term of the speaker's utility (Eq. 3) is shown throughout an interaction. When the speaker is initially uncertain about meanings (far left), the longer utterance u_1u_2 has higher expected informativity (center-left) and therefore higher utility (center-right) than the shorter utterances u_1 and u_2, despite its higher cost (far-right). As the speaker observes several successful interactions, they update their beliefs and become more confident about the meanings of the component lexical items u_1 and u_2. As a result, more efficient single-word utterances gradually gain in utility as cost begins to dominate the utility. On trial 5, u_1 is sampled, breaking the symmetry between utterances.

5). Consider a speaker who wants to refer to object ■. They expect their partner to be slightly more likely to interpret their language using a lexicon in which u_1 and u_2 apply to this object, due to their weak initial biases. However, there is still a reasonable chance ($p = 0.45$) that either u_1 or u_2 alone will be interpreted to mean ●, giving their partner false beliefs.

To see why our speaker model initially prefers the longer utterance u_1u_2 to hedge against this possibility, despite its higher production cost, consider the expected informativity of u_1u_2 under different possible lexicons. The possibility with highest probability is that both $\phi(u_1) = \phi(u_2) =$ ■ in the listener's lexicon ($p = 0.55^2 \approx 0.3$), in which case the listener will correctly identify ■ with high probability. The possibility that both $\phi(u_1) = \phi(u_2) =$ ● in the listener's lexicon is only $p = 0.45^2 \approx 0.2$, in which case the listener will erroneously select ●. In the mixed cases, where $\phi(u_1) =$ ●, $\phi(u_2) =$ ■ or $\phi(u_1) =$ ■, $\phi(u_2) =$ ● in the listener's lexicon ($p = 2 \cdot 0.45 * 0.55 \approx 0.5$), the utterance would be a interpreted as a contradiction and the listener would not change their prior beliefs. Because the speaker's informativity is defined using the log probability of the listener's belief, the utility of giving the listener a false belief (i.e. $\log(\epsilon)$) is significantly worse than simply being uninformative (i.e. $\log(0.5)$), and the longer utterance minimizes this harm.

Following the production of a conjunction, the speaker observes the listener's response (say, ■). This allows both agents to become more confident that the component utterances u_1 and u_2 mean ■ in their updated posterior over the listener's lexicon. This credit assignment to individual lexical items is a consequence of the compositional meaning of longer utterances in our simple grammar. The listener knows a speaker for whom either u_1 or u_2 individually means ■ would have been more likely to say u_1u_2 than a speaker for whom either component meant ●; and similarly for the speaker reasoning about possible listeners. Consequently, the probability of both mappings increases.

Fig. 6 shows the trajectories of internal components of the speaker utility as the interaction continues. We assume for illustrative purposes in this example that ■ continues to be the target on each trial and the same agent continues to be the speaker. As the posterior probability that individual primitive utterances u_1 and u_2 independently mean ■ increases (far left), the marginal gap in informativity between the conjunction and the shorter components gradually decreases (center left). As a consequence, production cost increasingly dominates the utility (center-right). After several trials of observing a successful listener response given the conjunction, the *informativity* of the two shorter utterances reaches parity with the conjunction but the cost makes

the shorter utterances more attractive (yielding a situation now similar to the outset of Simulation 1.1). Once the speaker samples one of the shorter utterances (e.g. u_1), the symmetry collapses and that utterance remains most probable in future rounds, allowing for a stable and efficient *ad hoc* convention. Thus, increasing efficiency is derived as a rational consequence of uncertainty and partner-specific inference about the listener's lexicon. For these simulations, we used $\alpha_S = \alpha_L = 8, w_c = 0.24, \beta = 0.8$ but the qualitative reduction effect is found over a range of different parameters (see Appendix Fig. A2).

Discussion

The simulations presented in this section aimed to establish a rational explanation for feedback-sensitive increases in efficiency over the course of *ad hoc* convention formation. Speakers initially hedge their descriptions under uncertainty about the lexical meanings their partner is using, but are able to get away with less costly components of those descriptions as their uncertainty decreases. Our explanation recalls classic observations about *hedges*, expressions like *sort of* or *like*, and morphemes like *-ish*, that explicitly mark provisionality, such as *a car, sort of silvery purple colored* (Fraser, 2010; Medlock & Briscoe, 2007). Brennan and Clark (1996) counted hedges across repetitions of a repeated reference game, finding a greater occurrence of hedges on early trials than later trials and a greater occurrence under more ambiguous contexts. While our model does not explicitly produce hedges, it is possible to understand this behavior as an explicit or implicit marker of the lexical uncertainty in our account. If hedges are explicit and intentional, this would require the speaker to reason about a listener that is to some degree aware of uncertainty or graded meanings. This is not the case with a classical truth-functional lexical representation, but is formalized naturally under soft semantics (see Appendix C).

We have already discussed why this phenomenon poses a challenge for the simple model-free reinforcement learning models in the literature — namely, that successful listener feedback would only reinforce long utterances with no mechanism for shortening them. This observation is not intended to rule out the entire family of reinforcement learning approaches, however. It is plausible that more sophisticated *model-based* reinforcement learning algorithms are flexible enough to account for the phenomenon. For instance, hierarchical architectures that appropriately incorporate compositionality or incrementality into the speaker's production model may be able to reinforce component parts of longer utterances in the shared history (e.g. Hawkins, Kwon, Sadigh, & Goodman, 2020). Still, such an approach would have more in common with our proposal than to the model-free heuristics in the existing literature. We return to this question with respect to scalability in the General Discussion.

Finally, the theory of reduction explored in this section is consistent with recent analyses of exactly *what* gets reduced in a large corpus of repeated reference games (Hawkins, Frank, & Goodman, 2020). These analyses found that entire modifying clauses are more likely to be dropped at once than expected by random corruption, and function words like determiners are mostly dropped as parts of larger noun phases or prepositional phrases rather than omitted on their own. In other words, speakers apparently begin by combining multiple descriptive 'units' and collapse to one of these 'units', rather than dropping words at random and assuming the listener can recover the intended longer utterance, as predicted by a noisy channel model. This theoretical claim is further supported by early hand-tagged analyses by Carroll (1980), which found that in three-quarters of transcripts from Krauss and Weinheimer (1964) the conventions that participants eventually converged upon were prominent in some syntactic construction at the beginning, often as a head noun that was initially modified or qualified by other information.

While our account explains these observations as a result of structure and heterogeneity in the lexical prior, it remains an open question of how to instantiate appropriately realistic priors in our computational model. Our simulation only considered two-word descriptions with homogenous uncertainty over the components, and is likely that the semantic components of real initial descriptions have more heterogeneous uncertainty: for example, the head noun may be chosen due to a higher prior probability of being understood by the listener than other components of the initial description, thus predicting asymmetries in reduction. Future work is needed to elicit these priors and evaluate predictions about more fine-grained patterns of reduction.

Phenomenon #2: Conventions gradually generalize to new partners in a social network

How do we make the inferential leap from *ad hoc* conventions formed through interaction with a single partner to *global* conventions expected to be shared throughout a community? Grounding collective convention formation in the individual learning mechanisms explored in the previous section requires an explicit *theory of generalization* capturing how people transfer what they have learned from one partner to the next.

One influential theory is that speakers simply ignore the identity of different partners and update a single monolithic representation after every interaction (Steels, 1995; Barr, 2004; Young, 2015). We call this a *complete-pooling* theory because data from each partner is collapsed into an undifferentiated pool of evidence (Gelman & Hill, 2006). Complete-pooling models have been remarkably successful at predicting collective behavior on networks, but have typically been evaluated only in settings where anonymity is enforced. For example, Centola and Baronchelli (2015) asked how large networks of participants coordinated on conventional names for novel faces. On each trial, participants were paired with a random neighbor but were not informed of that neighbor's identity, or the total number of different possible neighbors.

While complete-pooling may be appropriate for some everyday social interactions, such as coordinating with anonymous drivers on the highway, it is less tenable for everyday communicative settings. Knowledge about a partner's identity is both available and relevant for conversation (Eckert, 2012; Davidson, 1986). Partner-specificity thus poses clear problems for complete-pooling theories but can be easily explained by another simple model, where agents maintain separate expectations about meaning for each partner. We call this a *no-pooling* model (note that this no-pooling model was compared with a complete-pooling model in Smith et al., 2017). The problem with no-pooling is that agents are forced to start from scratch with each partner. Community-level expectations never get off the ground.

Our hierarchical *partial-pooling* account offers a compromise between these extremes. Unlike complete-pooling and no-pooling models, we propose that beliefs about language have hierarchical structure. That is, the meanings used by different partners are expected to be drawn from a shared community-wide distribution but are also allowed to differ from one another in systematic, partner-specific ways. This structure provides an inductive pathway for abstract population-level expectations to be distilled from partner-specific experience.

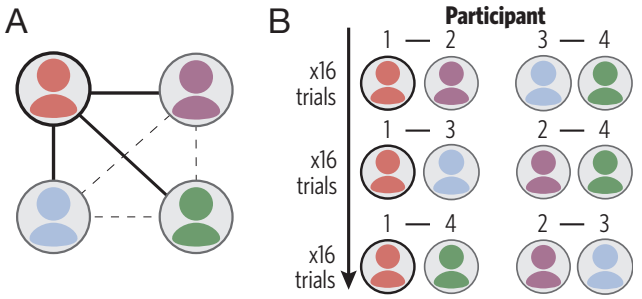

Figure 7. In our simulations and behavioral experiment, participants were (A) placed in fully-connected networks of 4, and (B) paired in a round-robin schedule of repeated reference games with each neighbor.

The key predictions distinguishing our model thus concern the pattern of generalization across partners. Experience with a single partner ought to be relatively uninformative about further partners, hence our partial-pooling account behaves much like a no-pooling model in predicting strong partner-specificity and discounting outliers (see Dautriche, Goupil, Smith, & Rabagliati, 2021, which explores this prediction in a developmental context). After interacting with multiple partners in a tight-knit community, however, speakers should become increasingly confident that labels are not simply idiosyncratic features of a particular partner's lexicon but are shared across the entire community, gradually transitioning to the behavior of a complete-pooling model. In this section, we test this novel prediction in a networked communication game and explicitly compare our model to pure complete-pooling and no-pooling variants.

Model predictions

We first compare the generalization behavior produced by each model by simulating the outcomes of interacting with multiple partners on a small network (see Fig. 7A). We used a round-robin scheme (Fig. 7B) to schedule four agents into a series of repeated reference games with their three neighbors, playing 8 successive trials with one partner before advancing to the next, for

a total of 24 trials. These reference games used a set of two objects $\{o_1, o_2\}$ and four utterances $\{u_1, u_2, u_3, u_4\}$ as in Simulation 1.2; agents were randomized to roles when assigned to a new partner and swap roles after each repetition block.

Unlike our previous simulations with a single partner, where hierarchical generalization was irrelevant, we must now specify the hyper-prior $P(\Theta)$ governing the overall distribution of *partners* (Eq. 4). Following Kemp et al. (2007), we extend the uniform categorical prior over objects to a hierarchical Dirichlet-Multinomial model (Gelman et al., 2014), where the categorical prior over the partner-specific meaning of u, $P(\phi_k(u_i) = o_j)$, is not uniform, but given by a parameter Θ that is shared across the entire population. Because Θ is a vector of probabilities that must sum to 1, we assume it is drawn from a Dirichlet prior:

$$\begin{aligned} \phi_k(u) &\sim \text{Categorical}(\Theta) \\ \Theta &\sim \text{Dirichlet}(\boldsymbol{\alpha}) \end{aligned}$$

where $\boldsymbol{\alpha}$ is the concentration parameter encoding the agent's beliefs, or "over-hypotheses" about both the central tendency and the variability of lexicons in the population. The relative values of the entries of $\boldsymbol{\alpha}$ correspond to inductive biases regarding the central tendency of lexicons while the absolute magnitude of $\boldsymbol{\alpha}$ roughly corresponds to prior beliefs about the spread, where larger magnitudes correspond to more concentrated probability distributions. We assume the agent also has uncertainty about these population-level properties with a hyper-prior on α (see Cowans, 2004) roughly corresponding to the weak initial preferences we used in our previous simulations:

$$\boldsymbol{\alpha} \sim \begin{cases} \text{Dirichlet}(2,3) & \text{if} \quad u \in \{u_1, u_2\} \\ \text{Dirichlet}(3,2) & \text{if} \quad u \in \{u_3, u_4\} \end{cases}$$

We may then define the no-pooling and complete-pooling models by lesioning this shared structure in different ways. The no-pooling model assumes an independent (Θ_k, α_k) for every partner, rather than sharing a single population-level parameter. Conversely, the complete-pooling model assumes a single, shared ϕ rather than allowing different values ϕ_k for different partners. We simulated 48 networks for each model, setting $\alpha_S = \alpha_L = 4$, $w_C = .24$ (see Fig. A3 in the Appendix for an exploration of other parameters).

Network convergence. Because all agents are simultaneously making inferences about the others, the network as a whole faces a coordination problem. For example, in the first block, agents 1 and 2 may coordinate on using u_1 to refer to o_1 while agent 3 and 4 coordinate on using u_2. Once they swap partners, they must negotiate this potential mismatch in usage. How does the network as a whole manage to coordinate?

We measured alignment by examining the intersection of utterances produced by speakers: if two agents produced overlapping utterances to refer to a given target (i.e. a non-empty intersection), we assign a 1, otherwise we assign a 0. At the beginning and end of each interaction, we calculated alignment between currently interacting agents (i.e. *within* a dyad) and those who were not interacting (i.e. *across* dyads), averaging across target objects. Alignment across dyads was initially near chance, reflecting the arbitrariness of whether speakers reduce to u_1 or u_2. Under a no-pooling model (Fig. 8C, third row), subsequent blocks remain near chance, as conventions need to be renegotiated from scratch. Under a complete-pooling model (Fig. 8C, second row), agents persist with mis-calibrated expectations learned from previous partners rather than adapting to their new partner, and *within-dyad* alignment deteriorates. By contrast, under our partial-pooling model, alignment across dyads increases without affecting alignment within dyads, suggesting that hierarchical inference leads to emergent consensus (Fig. 8C, top row).

Speaker utterance length across partners. Next, we examined our model's predictions about how a *speaker*'s referring expressions will change with successive listeners. While it has been frequently observed that messages reduce in length across repetitions with a single partner (Krauss & Weinheimer, 1964) and sharply revert back to longer utterances when a new partner is introduced (Wilkes-Gibbs & Clark, 1992), the key prediction distinguishing our model concerns behavior across subsequent partner boundaries. Complete-pooling accounts predict no reversion in number of words when a new partner is introduced (Fig. 8B, second row). No-pooling accounts predict that roughly the same initial description length will reoccur with every subsequent interlocutor (Fig. 8B, third row).

Here we show that a partial pooling account predicts

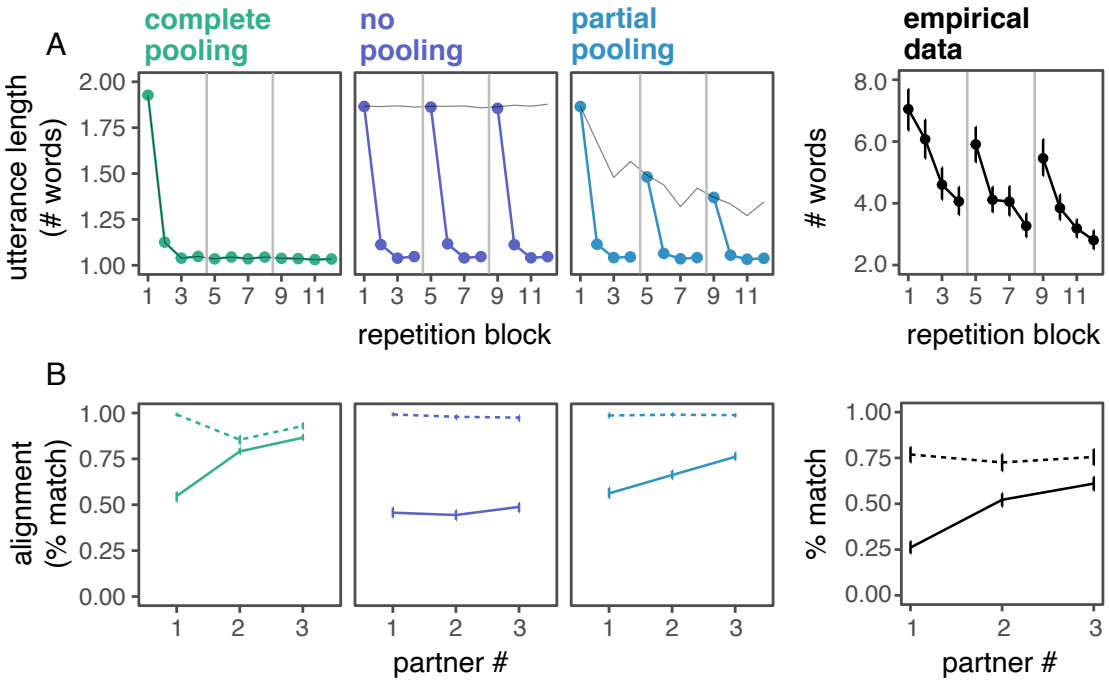

Figure 8. Simulation results and empirical data for (A) speaker reduction, and (B) network convergence across three partners. Vertical lines mark boundaries where new partners were introduced, and thin grey line represents the utterance the model would produce for a novel partner at each point in time.

a more complex pattern of generalization. First, unlike the complete-pooling model, we find that the partial-pooling speaker model reverts or jumps back to a longer description at the first partner swap. This reversion is due to ambiguity about whether the behavior of the first partner was idiosyncratic or attributable to community-level conventions. In the absence of data from other partners, a partner-specific explanation is more parsimonious. Second, unlike a no-pooling model, after interacting with several partners, the model becomes more confident that one of the short labels is shared across the entire community, and is correspondingly more likely to begin a new interaction with it (Fig. 8B, top row).

It is possible, however, that these two predictions only distinguish our partial-pooling models at a few parameter values; the no-pooling and complete-pooling could produce these qualitative effects elsewhere in parameter space. To conduct a more systematic model comparison, then, we simulated 10 networks in each cell of a large grid manipulating the the optimality parameters α_S, α_L, the cost parameter w_C, and the memory discounting parameter β. We computed a "reversion" statistic (the magnitude of the change in $P(u_1u_2)$ immediately after a partner swap) and a "generalization" statistic (the magnitude of the change in $P(u_1u_2)$ from the initial trial with the agent's first partner to the initial trial with the final partner) and conducted single-sample t-tests at each parameter value to compare these statistics with what would be expected due to random variation. We found that only the partial-pooling model consistently makes both predictions across a broad regime. The complete-pooling model fails to predict reversion nearly everywhere while the no-pooling model fails to predict generalization nearly everywhere. Detailed results are shown in Fig. A4 in the Appendix.

Behavioral experiment

To evaluate the predictions derived in our simulations, we designed a natural-language communication experiment following roughly the same network design as our simulations. That is, instead of anonymizing partners, as in many previous empirical studies of convention formation (e.g. Centola & Baronchelli, 2015), we divided the experiment into blocks of extended dyadic interactions with stable, identifiable part-

ners (see Fay et al., 2010; Garrod & Doherty, 1994, for similar designs). Each block was a full repeated reference game, where participants had to coordinate on *ad hoc* conventions for how to refer to novel objects with their partner. Our partial-pooling model predicted that these conventions will partially reset at partner boundaries, but agents should be increasingly willing to transfer expectations from one partner to another.

Participants. We recruited 92 participants from Amazon Mechanical Turk to play a series of interactive, natural-language reference games using the framework described in Hawkins (2015).

Stimuli and procedure. Each participant was randomly assigned to one of 23 fully-connected networks with three other participants as their 'neighbors' (Fig. 7A). Each network was then randomly assigned one of three distinct "contexts" containing abstract tangram stimuli taken from (Clark & Wilkes-Gibbs, 1986). The experiment was structured into a series of three repeated reference games with different partners, using these same four stimuli as referents. Partner pairings were determined by a round-robin schedule (Fig. 7B). The trial sequence for each reference game was composed of four repetition blocks, where each target appeared once per block. Participants were randomly assigned to speaker and listener roles and swapped roles on each block. After completing sixteen trials with one partner, participants were introduced to their next partner and asked to play the game again. This process repeated until each participant had partnered with all three neighbors. Because some pairs within the network took longer than others, we sent participants to a temporary waiting room if their next partner was not ready.

Each trial proceeded as follows. First, one of the four tangrams in the context was highlighted as the *target object* for the speaker. They were instructed to use a chatbox to communicate the identity of this object to their partner, the listener (see Fig. 7C). The two participants could engage freely in dialogue through the chatbox but the listener must ultimately make a selection from the array. Finally, both participants in a pair were given full feedback on each trial about their partner's choice and received bonus payment for each correct response. The order of the stimuli on the screen was randomized on every trial to prevent the use of spatial cues (e.g. 'the one on the left'). The display also contained an avatar for the current partner representing different partners with different colors as shown in Fig. 7 to emphasize that they were speaking to the same partner for an extended period. On the waiting screen between partners, participants were shown the avatars of their previous partner and upcoming partner and told that they were about to interact with a new partner.

Results

We evaluated participants' generalization behavior on the same three metrics we used in our simulations: accuracy, utterance length, and network convergence.

Network convergence. First, we examine the *content* of conventions and evaluate the extent to which alignment increased across the network over the three partner swaps. Specifically, we extend the same measure of alignment used in our simulations to natural language data by examining whether the intersection of words produced by different speakers was non-empty. We excluded a list of common stop words (e.g. 'the', 'both') to focus on the core conceptual content. While this pure overlap measure provides a relatively weak notion of similarity, a more continuous measure based on the *size* of the intersection or the string edit distance yielded similar results.

As in our simulation, the main comparison of interest was between currently interacting participants and participants who are not interacting: we predicted that within-pair alignment should stay consistently high while (tacit) alignment between non-interacting pairs will increase. We thus constructed a mixed-effects logistic regression including fixed effects of pair type (within vs. across), partner number, and their interaction. We included random intercepts at the tangram level and maximal random effects at the network level (i.e. intercept, both main effects, and the interaction). As predicted, we found a significant interaction ($b = -0.85, z = -5.69, p < 0.001$; see Fig. 8C, bottom row). Although different pairs in a network may initially use different labels, these labels begin to align over subsequent interactions.

Speaker utterance length. Now we are in a position to evaluate the central prediction of our model. Our partial pooling model predicts (1) gains in efficiency within interactions with each partner and (2) reversions to longer utterances at partner boundaries, but (3) gradual shortening of the initial utterance chosen with successive partners. As a measure of effi-

ciency, we calculated the raw length (in words) of the utterance produced on each trial. Because the distribution of utterance lengths is heavy-tailed, we log-transformed these values. To test the first prediction, we constructed a linear mixed-effects regression predicting trial-level speaker utterance length (see Fig. 8B, bottom row). We included a fixed effect of repetition block within partner (1, 2, 3, 4), along with random intercepts and slopes for each participant and each tangram. We found that speakers reduced utterance length significantly over successive interactions with each individual partner, $b = -0.19, t(34) = -9.88, p < 0.001$.

To test the extent to which speakers revert to longer utterances at partner boundaries, we constructed another regression model. We coded the repetition blocks immediately before and after each partner swap, and included it as a categorical fixed effect. Because partner roles were randomized for each game, the same participant did not always serve as listener in both blocks, so in addition to tangram-level intercepts, we included random slopes and intercepts at the *network* level (instead of the participant level). As predicted, we found that utterance length increased significantly at the two partner swaps, $b = 0.43, t(22) = 4.4, p < 0.001$.

Finally, to test whether efficiency improves for the *very first* interaction with each new partner, before observing any partner-specific information, we examined the simple effect of partner number at the trials immediately after the partner swap (i.e. $t = \{1, 5, 9\}$). We found that participants gradually decreased the length of their initial descriptions with each new partner in their network, $b = -0.2, t(516.5) = -6.07, p < 0.001$ (see Fig. 8B, bottom row), suggesting that speakers are bringing increasingly well-calibrated expectations into interactions with novel neighbors. The partial-pooling model is the only model predicting all three of these effects.

Discussion

Our partial pooling model, which follows from general principles of hierarchical Bayesian inference, suggests that conventions represent the shared structure that agents "abstract away" from partner-specific learning. In this section, we evaluated the extent to which our partial pooling model captured human generalization behavior in a natural-language communication experiment on small networks. Unlike complete-pooling accounts, our model allows for partner-specific common ground to override community-wide expectations given sufficient experience with a partner, or in the absence of strong conventions. Unlike no-pooling accounts, it results in networks that converge on more efficient and accurate expectations about novel partners.

While it is not easily classified into either the complete-pooling or no-pooling classes, it is also not straightforward for the priming mechanisms proposed by *interactive alignment* accounts to explain these patterns of partner-specificity without being augmented with additional social information. If a particular semantic representation has been activated due to precedent in the preceding dialogue, then the identity of the speaker should not in principle alter its continued influence (Brennan & Hanna, 2009). More sophisticated hierarchical memory retrieval accounts that represent different partners as different *contexts* (e.g Polyn, Norman, & Kahana, 2009) may be consistent with partner-specificity, but evoking such an account presupposes that social information like partner identity is already a salient and relevant feature of the communicative environment and thus no longer relies purely on "egocentric" priming and activation mechanisms. Indeed, a process-level account assuming socially-aware context reinstatement for partner-specific episodic memories, and slower consolidation of shared features into population-level expectations, may be one possible candidate for realizing our computational-level model (see General Discussion for more on process-level evidence).

Phenomenon #3: Conventions are shaped by communicative context

In the previous two sections, we examined a mechanism for rapid, partner-specific learning that allows agents to form stable but arbitrary *ad hoc* conventions with partners and gradually generalize them to their entire community. The final phenomenon we consider is the way that *ad hoc* conventions are shaped by the communicative context in which they form. This phenomenon is most immediately motivated by recent behavioral results finding that more informative words in the local context are significantly more likely to become conventionalized (Hawkins, Frank, & Goodman, 2020). However, our broader theoretical aim is to suggest that *diachronic* patterns in the long-term evolution of a community's lexical conventions, as highlighted by the Op-

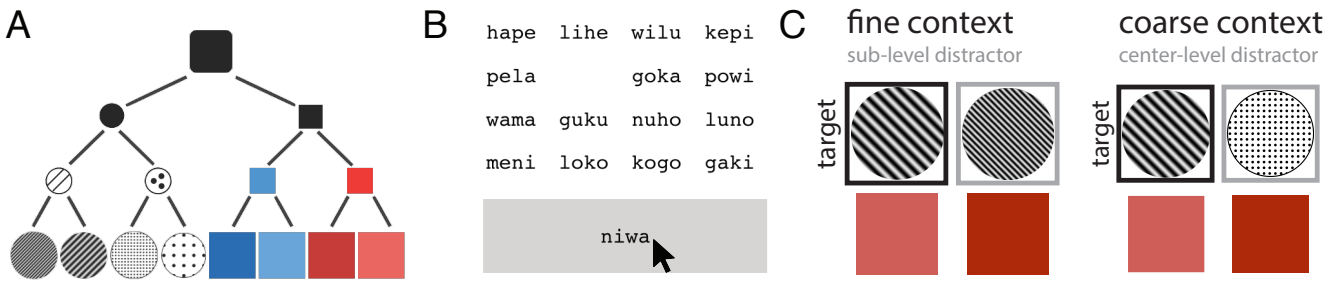

Figure 9. Domain for context-sensitivity. (A) Targets are related to one another in a conceptual taxonomy. (B) Speakers choose between labels, where the label "niwa" has been selected. (C) Examples of *fine* and *coarse* contexts. In the *fine* context, the target (marked in black) must be disambiguated from a distractor (marked in grey) at the same subordinate-level branch of the taxonomy. In the *coarse* context, the closest distractor belongs to a different branch of the center-level of the taxonomy (i.e. a spotted circle) such that disambiguation at the sub-ordinate level is not required.

timal Semantic Expressivity (OSE) hypothesis (Frank, 2017), may be explained as a result of the *synchronic* processes at play when individual dyads coordinate on *ad hoc* meanings.

Briefly, when there is already a strong existing convention that is expected to be shared across the community, our model predicts that speakers will use it. New *ad hoc* conventions arise precisely to fill *gaps* in existing population-level conventions, to handle new situations where existing conventions are not sufficient to accurately and efficiently make the distinctions that are required in the current context. A corollary of this prediction[7] is that *ad hoc* conventions may only shift to expectations at the population level (and ultimately to population-level convergence) when those distinctions are consistently relevant across interactions with different partners. For example, while most English speakers have the basic-level word "tree" in their lexicon, along with a handful of subordinate-level words like "maple" or "fir," we typically do not have conventionalized labels exclusively referring to each individual tree in our yards – we are rarely required to refer to individual trees. Meanwhile, we *do* often have shared conventions (i.e. proper nouns) for individual people and places that a community regularly encounters and needs to distinguish among. Indeed, this logic may explain why a handful of particularly notable trees do have conventionalized names, such as the Fortingall Yew, the Cedars of God, and General Sherman, the giant sequoia.

As a first step toward explaining these diachronic patterns in *which* conventions form, we aim to establish in this section that our model allows a single dyad's *ad hoc* conventions to be shaped by communicative context over short timescales. Specifically, our model predicts that people will form conventions at the highest level of abstraction that is able to satisfy their communicative needs. That is, when the local environment imposes a communicative need to refer to particular *ad hoc* concepts (e.g. describing a particular tree that needs to be cut down), communicative partners are able to coordinate on efficient lexical conventions for successfully doing so (e.g. "the mossy one").

First, we show that context-sensitivity naturally emerges from our model, as a downstream consequence of recursive pragmatic reasoning. We then empirically evaluate this account by manipulating which distinctions are relevant in an artificial-language repeated reference game building on Winters et al. (2014, 2018), allowing us to observe the emergence of *ad hoc* conventions from scratch. In both the empirical data and our model simulations, we find that conventions come to reflect the distinctions that are functionally relevant for communicative success and that pragmatic reasoning is needed for these effects to arise.

Model predictions

To evaluate the impact of context on convention formation, we require a different task than we used in

[7]This follows by induction from the hierarchical generalization mechanisms evaluated for **P2**, which provide the pathway by which *ad hoc* conventions become adopted by a larger community over longer time scales. Many *ad hoc* conventions never generalize to the full language community simply because the contexts where they are needed are rare and must continued to be re-negotiated with subsequent partners on an *ad hoc* bases.

the previous sections. Those tasks, like most reference games in the literature on convention formation, used a discrete set of unrelated objects in a fixed context, $\{o_1, \ldots, o_k\}$. In real referential contexts, however, targets are embedded in larger conceptual taxonomies, where some objects are more similar than others (Bruner, Goodnow, & Austin, 1956; Collins & Quillian, 1969; Xu & Tenenbaum, 2007). Here, we therefore consider a space of objects embedded in a three-level stimulus hierarchy with shape at the topmost level, color/texture at the intermediate levels, and frequency/intensity at the finest levels (see Fig. 9A). While we will use the full stimulus set in our empirical study, it is sufficient for our simulations to consider just one of the branches (i.e. just the squares).

We populate the space of possible utterance meanings $P(\phi)$ with 4 meanings at the sub-ordinate level (one for each individual object, e.g. $\phi(u)$ = "light blue square"), 2 meanings at the center-level (e.g. $\phi(u)$ = "blue square"), and 1 meanings at the super-ordinate level (e.g. $\phi(u)$ = "square"). We then populate the utterance space with 8 single-word labels (Fig. 9B) and also allow for a "null" meaning with an empty extension to account for the possibility that some utterances are not needed, allowing the agent to effectively remove utterances from their vocabulary.

Additionally, in real environments, speakers do not have the advantage of a fixed context; the relevant distinctions change from moment to moment as different subsets of objects are in context at different times. This property poses a challenge for models of convention formation because the relevant distinctions cannot be determined from a single context, they must be abstracted over time. We therefore only displayed four of the eight objects in the context on a given trial. Distractors could differ from the target at various levels of the hierarchy, creating different types of contexts defined by the finest distinction that had to be drawn (e.g. Fig. 9C).

Critically, we manipulated the prevalence of different kinds of contexts, controlling how often participants are required to make certain distinctions to succeed at the task. In the *fine* condition, every context contained a subordinate distractor, requiring fine low-level distinctions to be drawn. In the *coarse* condition, contexts never contained subordinate distractors, only distractors that differed at the central level of the hierarchy (e.g. a blue square when the target is a red square). For comparison, we also include a *mixed* condition, where half of the targets appeared in *fine* contexts with subordinate distractors and the other half appeared in *coarse* contexts without them.

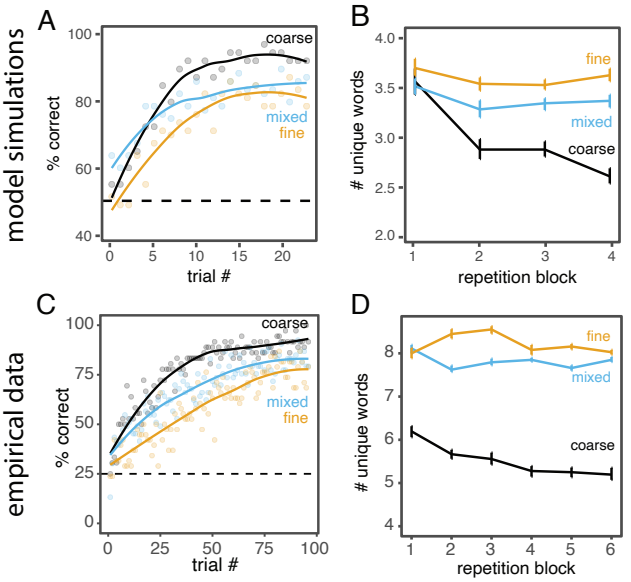

Figure 10. Comparison of simulation results to empirical data. (A) Agents in our simulation learn to coordinate on a successful communication system, but converge faster in the coarse condition than the fine condition. (B) The number of unique words used by agents in each repetition block increased in the fine condition but stayed roughly constant in the coarse condition. (C-D) The same metrics computed on our empirical data, qualitatively matching the patterns observed in the simulations. Each point is the mean proportion of correct responses by listeners; curves are nonparametric fits.

We constructed the trial sequence identically for the three conditions. On each trial, we sampled one of the four objects to be the target, ensuring that no target appeared more than once in a row. Then we sampled a distractor according to the constraints of the condition. As before, the agents swap roles on each trial, and we run 50 distinct trajectories with parameter settings of $\alpha_L = 8, \alpha_S = 8$ and memory discounting parameter of $\beta = 0.6$.

Partners successfully learn to communicate. First, we compare the model's learning curves across context conditions (Fig. 10A). We focus on the *coarse*

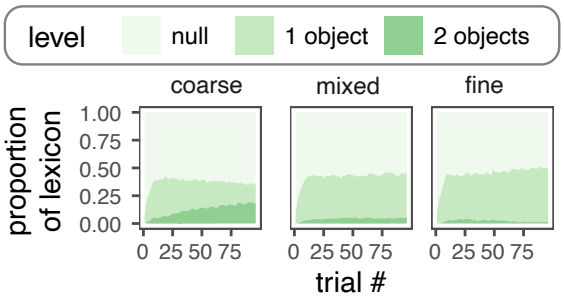

Figure 11. Dynamics of lexical beliefs over time in model simulations. Regions represent the average proportion of words at each level of generality in an agent's beliefs about the lexicon. In the coarse condition, agents initially assume subordinate terms but gradually abstract away to a smaller number of more general terms; in the fine and mixed conditions, however, agents become more confident of subordinate terms.

and *fine* conditions for simplicity, since this single comparison captures the core phenomena of interest. In a mixed-effects logic regression, we find that communicative accuracy steadily improves over time across all conditions, $b = 0.72, z = 16.9, p < 0.001$. However, accuracy also differed across conditions: adding a main effect of condition significantly improves model fit, $\chi^2(2) = 9.6, p = 0.008$. Accuracy is significantly higher in the coarse condition than the fine condition $b = -0.71, z = 9.3, p < 0.001$ and marginally higher than the mixed condition.

Lexical conventions are shaped by context. As an initial marker of context sensitivity, we examine the effective vocabulary sizes used by speakers in each condition. We operationalized this measure by counting the total number of unique words produced within each repetition block. This measure takes a value of 8 when a different word is consistently used for every object, and a value of 1 when exactly the same word is used for every object. In an mixed-effects regression model including intercepts and random effects of trial number for each simulated trajectory, we find an overall main effect of condition, with agents in the fine condition using significantly fewer words across all repetition blocks ($m = 4.7$ in *coarse*, $m = 6.5$ in *fine,* $t = 4.5, p < 0.001$). However, we also found a significant interaction: the effective vocabulary size gradually increased over time in the fine condition, while it stayed roughly constant in in the coarse condition, $b = 0.18, t = 8.1, p < 0.001$, see Fig. 10B.

Next, we examine more closely the emergence of terms at different levels of generality. We have access not only to the signaling behavior of our simulated agents, but also their internal *beliefs* about their partner's lexicon, which allows us to directly examine the evolution of these beliefs from the beginning of the interaction. At each time point in each game, we take the single meaning with highest probability for each word. In Fig. 11, we show the proportion of words with meanings at each level of generality, collapsing across all games in each condition.

Qualitatively, we observe that agents begin by assuming null meanings (i.e. with an effectively empty vocabulary) but quickly begin assigning meanings to words based on their partner's usage. In both conditions, basic-level meanings and subordinate-meanings are equally consistent with the initial data, but the simplicity prior prefers smaller effective vocabulary sizes where each word has smaller extensions. After the first repetition block, however, agents in the coarse condition begin pruning out some of the subordinate-level terms and become increasingly confident of basic-level meanings. Agents in the fine condition become even more confident of subordinate level meanings.

By the final trial, the proportion of basic-level vs. subordinate-level terms is significantly different across the coarse and fine conditions. Only 9% of words had subordinate-level meanings (green) in the coarse condition, compared with 79% in the fine condition, $\chi^2(1) = 436, p < 0.001$. At the same time, 45% of words had basic-level meanings (blue) in the coarse condition, compared with only 8% in the fine condition, $\chi^2(1) = 136, p < 0.001$. The remaining words in each condition were assigned the 'null' meaning (red), consistent with an overall smaller effective vocabulary size in the coarse condition. The diverging conventions across contexts are driven by Gricean expectations: because the speaker is assumed to be informative, only lexicons distinguishing between subordinate level objects can explain the speaker's behavior in the *fine* condition.

Experimental methods

In this section, we evaluate our model's qualitative predictions about the effect of context on convention formation using an interactive behavioral experiment

closely matched to our simulations. We use a between-subjects design where pairs of participants are assigned to different communicative contexts and test the extent to which they converge on meaningfully different conventions.

Participants. We recruited 278 participants from Amazon Mechanical Turk to play an interactive, multiplayer game.[8].

Procedure & Stimuli. Participants were paired over the web and placed in a shared environment containing an array of objects (Fig. 9A) and a 'chatbox' to choose utterances from a fixed vocabulary by clicking-and-dragging (Fig. 9B). On each trial, one player (the 'speaker') was privately shown a highlighted target object and allowed to send a single word to communicate the identity of this object to their partner (the 'listener'), who subsequently made a selection from the array. Players were given full feedback, swapped roles each trial, and both received bonus payment for each correct response.

We randomly generated distinct arrays of 16 utterances for each pair of participants (more than our model, which was restricted by computational complexity). These utterances were created by stringing together consonant-vowel pairs into pronounceable 2-syllable words to reduce the cognitive load of remembering previous labels (see Fig. 9B) These arrays were held constant across trials.

To match our model as closely as possible, pairs were assigned one of the same sequences of trials that we constructed for our simulations. In addition to behavioral responses collected over the course of the game, we designed a post-test to explicitly probe players' final lexica. For all sixteen words, we asked players to select all objects that a word can refer to (if any), and for each object, we asked players to select all words that can refer to it (if any). This bidirectional measure allowed us to check the internal validity of the lexica reported. Pairs were randomly assigned to one of three different conditions, yielding $n = 36$ dyads in the *coarse* condition, $n = 38$ in the *fine* condition, and $n = 53$ in the *mixed* condition after excluding participants who disconnected before completion.

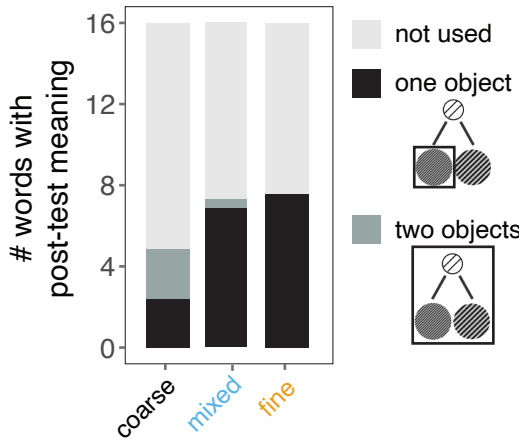

Figure 12. Different lexicons emerge in different contexts. Mean number of words, out of a word bank of 16 words, that human participants reported giving more specific meanings (black; applying to 1 object) or less specific meanings (dark grey; applying to 2 objects) in the post-test.

Behavioral results

Partners successfully learn to communicate. Although participants in all conditions began with no common basis for label meanings, performing near chance on the first trial (proportion correct = 0.19, 95% CI = $[0.13, 0.27]$), most pairs were nonetheless able to coordinate on a successful communication system over repeated interaction (see Fig. 10C). A mixed-effects logistic regression on listener responses with trial number as a fixed effect, and including by-pair random slopes and intercepts, showed a significant improvement in accuracy overall, $z = 14.4, p < 0.001$. Accuracy also differed significantly *across* conditions: adding an additional main effect of condition to our logistic model provided a significantly better fit, $\chi^2(2) = 10.8$, $p = 0.004$. Qualitatively, the *coarse* condition was easiest for participants, the *fine* condition was hardest, and the *mixed* condition was in between. These effects track the most important qualitative feature of our simulations – our artificial agents were also able to successfully coordinate in both conditions, and did so more easily in the coarse condition than the fine condition. However, we found that the speed of coordination in the mixed and fine conditions was larger than predicted in our simulations. The additional difficulty participants' experienced in the *fine* condition may be due to addi-

[8]This experiment was pre-registered at `https://osf.io/2hkjc/`. All statistical tests in mixed-effects models reported in this section use degrees of freedom based on the Satterthwaite approximation (Luke, 2017).

tional motivational constraints, memory constraints, or other factors not captured in our model.

Validating post-test responses. Before examining post-test responses, we validate their internal consistency. For each participant, we counted the number of mismatches between the two directions of the lexicon question (e.g. if they clicked the word 'mawa' when we showed them one of the blue squares, but failed to click that same blue square when we showed 'mawa'). In general, participants were highly consistent: out of 128 cells in the lexicon matrix (16 words × 8 objects), the median number of mismatches was 2 (98% agreement), though the distribution has a long tail (mean = 7.3). We therefore conservatively take a participant's final lexicon to be the *intersection* of their word-to-object and object-to-word responses for the subsequent analyses.

Contextual pressures shape the lexicon. We predicted that in contexts regularly requiring speakers to make fine distinctions among objects at subordinate levels of the hierarchy, we would find lexicalization of specific terms for each object (indeed, a one-to-one mapping may be the most obvious solution in a task with only 8 objects). Conversely, when no such distinctions were required, we expected participants to adaptively conventionalize more general terms that could be reused across different contexts. One coarse signature of this prediction lies in the *compression* of the resulting lexicon: less specific conventions should allow participants to achieve the same communicate accuracy with a smaller vocabulary.

To test this prediction, we operationalize vocabulary size as the number of words in each participant's reported lexicon in the post-test (i.e. the words for which they marked at least one object in the post-test in an internally consistent way). We then conducted a mixed-effects regression predicting each individual's vocabulary size as a function of dummy-coded condition factors, with random intercepts for each game. We found that participants in the *coarse* condition reported significantly smaller, simpler lexica ($m = 4.9$ words) than participants in the *mixed* ($m = 7.1, t(110.8) = 7.6, p < 0.001$) and *fine* condition ($m = 6.9, t(109.3) = 6.4, p < 0.001$; see Fig. 12).

What allowed participants in the *coarse* condition get away with fewer words in their lexicon while maintaining high accuracy? We hypothesized that each word had a larger extension size. To test this hypothesis, we counted the numbers of 'specific' terms (e.g. words that refer to only one object) and more 'general' terms (e.g. words that refer to two objects) in the post-test. We found that the likelihood of lexicalizing more general terms differed systematically across conditions. Participants in the *coarse* condition reported significantly more general terms ($m = 2.3$) than in the *fine* ($m = 0.24, t(121.2) = 8.5, p < 0.001$) or *mixed* ($m = 0.65, t(122.7) = 7.3, p < 0.001$) conditions, where lexicons contained almost exclusively specific terms. Using the raw extension size of each word as the dependent variable instead of counts yielded similar results. Indeed, the modal system in the fine condition was exactly eight specific terms with no more general terms, and the modal system in the coarse condition was exactly four general terms (red, blue, striped, spotted) with no specific terms. However, many individual participants reported a mixture of terms at different levels of generality (see Appendix Fig. A6).

Finally, how did these lexica emerge over the course of interaction? We use the same measure of unique words produced in each repetition block that we used in our simulations (Fig. 10D). We constructed a mixed-effects regression model predicting the effective vocabulary size, including fixed effects of condition and repetition block, and random intercepts and effects of repetition block for each dyad. As in the post-test reports, we found an overall main effect of condition, with participants in the coarse condition using significantly fewer words across all repetition blocks: $m = 4.9$ in coarse, compared to $m = 5.8, t(124) = 6.1, p < 0.001$ in mixed and $m = 6.8, t(124) = 11.4, p < 0.001$ in fine. Critically, however, we also found a significant interaction between block and condition. The effective vocabulary size gradually increased over time in the fine condition but remained roughly constant in the coarse condition, $b = 0.12, t = 4.5, p < 0.001$, see Fig. 10D. This interaction, where participants initially attempt to reuse the same terms across targets in the fine condition, is consistent with a gradual differentiation based on communicative need.

Discussion

There is abundant evidence that languages adapt to the needs of their users. Our model provides a cognitive account of how people coordinate on *ad hoc* linguistic conventions that suit their immediate needs. In this sec-

tion, we evaluated predictions about context-sensitivity using new data from a real-time communication task. When combined with the generalization mechanisms explored in the previous section, such rapid learning within dyadic interactions may be a powerful contributor allowing languages to adapt at the population-level over longer time scales.

Previous studies of convention formation have addressed context-sensitivity in different ways. In some common settings, there is no explicit representation of context at all, as in the task known as the "Naming Game" where agents coordinate on names for objects in isolation (Steels, 2012; Baronchelli, Loreto, & Steels, 2008). In other settings, communication is situated in a referential context, but this context is held constant, as in Lewis signaling games (Lewis, 1969) where agents must distinguish between a fixed set of world states (Skyrms, 2010; Bruner, O'Connor, Rubin, & Huttegger, 2014). Finally, in the Discrimination Game (Steels & Belpaeme, 2005; Baronchelli, Gong, Puglisi, & Loreto, 2010), contexts are randomly generated on each trial, but have not been manipulated to assess context-sensitivity of the convention formation process.

In other words, context-sensitivity has typically been implicit in existing models. Models using simple update rules have accounted for referential context with a *lateral inhibition* heuristic used by both the speaker and listener agents (Franke & Jäger, 2012; Steels & Belpaeme, 2005). If communication is successful, the connection strength between the label and object is not only increased, the connection between the label and competing objects (and, similarly, between the object and competing labels) is explicitly *decreased* by a corresponding amount. This lateral inhibition heuristic is functionally similar to our pragmatic reasoning mechanism, in terms of allowing the agent to learn from negative evidence (i.e. the speaker's choice *not* to use a word, or the listener's choice *not* to pick an object). Under our inferential framework, however, this property emerges as a natural consequence of well-established Gricean principles of pragmatic reasoning rather than as a heuristic.

General Discussion

In this paper, we considered the computational challenges posed by communication in a variable and non-stationary landscape of meaning. We advanced a hierarchical Bayesian approach in which agents continually adapt their beliefs about the form-meaning mapping used by each partner, in turn. We formalized this approach by integrating three core cognitive capacities in a probabilistic framework: representing initial uncertainty about what a partner thinks words mean **(C1)**, partner-specific adaptation based on observations of language use in context **(C2)**, and hierarchical structure for graded generalization to new partners **(C3)**. This unified model resolves several puzzles that have posed challenges from prior models of coordination and convention formation: why referring expressions shorten over repeated interactions with the same partner **(P1)**, how partner-specific common ground may coexist with the emergence of conventions at the population level **(P2)**, and how context shapes which conventions emerge **(P3)**.

We conclude by raising five broader questions that follow from the theoretical perspective we have advanced, each suggesting pathways for future work: (1) are the mechanisms underlying *ad hoc* convention formation in adults the same as those underlying language acquisition in children? (2) to what extent do these mechanisms depend on the communication modality? (3) what kinds of feedback are used to coordinate on conventions? (4) what kinds of social structure are reflected in the hierarchical prior, and (5) which representations are involved in adaptation at a process-level?

Continuity of language learning across development

There is a close mathematical relationship between our model of convention formation and recent probabilistic models of word learning in developmental science (e.g. Xu & Tenenbaum, 2007; Frank et al., 2009). This similarity suggests an intriguing conjecture that the continual-learning mechanisms adults use to rapidly coordinate on partner-specific conventions may be the same as those supporting lexical acquisition in children. In other words, people may never stop learning language; they may simply develop stronger and better-calibrated beliefs about their community's conventions, which apply to a broader swath of communicative scenarios. In this section, we discuss three possible implications of viewing language development in terms of social coordination and convention.

First, common paradigms for cross-situational learn-

ing typically assume a fixed speaker and focus on variability across referential contexts (Siskind, 1996; Regier, 2005; Smith, Suanda, & Yu, 2014; Yurovsky & Frank, 2015). Yet, as we have argued, variability in how words are used by different *partners* may pose an equally challenging problem. While there has been limited work on the social axis of generalization, it is increasingly apparent that children are able to track *who* produced the words they are learning and use this information to determine whether word meanings should generalize not just to other contexts but also to other speakers. For example, young children may limit the generalizability of observations from speakers who use language in idiosyncratic ways, such as a speaker who call a ball a "dog" (Koenig & Woodward, 2010; Luchkina, Sobel, & Morgan, 2018), and even retrospectively update beliefs about earlier words after observing such idiosyncracies (Dautriche et al., 2021).

This discounting of idiosyncratic speakers may be understood as an instance of the same inductive problem that convention formation poses for adults in **P2**. Unlike complete-pooling models, which predict that all observations should be taken as equally informative about the community's conventions, our hierarchical model predicts that children should be able to explain away "outliers" without their community-level expectations being disrupted. Indeed, a novel prediction generated by our account is that children should be able to accommodate idiosyncratic language *within* extended interaction with the same speaker (e.g. continue to pretend the ball is called "dog," given partner-specific common ground) while also limiting generalization of that convention *across* other speakers.

Second, we have emphasized the importance of representing lexical *uncertainty*, capturing expected variability in the population beyond the point estimates assumed by traditional lexical representations. But how do children calibrate their lexical uncertainty? One possibility assigns a key role to the number of distinct speakers in a child's environment, by analogy to the literature on talker variability (Creel & Bregman, 2011; Clopper & Pisoni, 2004). Exposure to fewer partners may result in weaker (e.g. Lev-Ari, 2017) or miscalibrated priors for meanings. If an idiosyncratic construction is over-represented in the child's environment, they may later be surprised to find that it was specific to their household's lexicon and not shared by the broader community (see Clark, 2009, Chap. 6). Conversely, however, hierarchical inference predicts a blessing of abstraction (Goodman et al., 2011): under certain conditions, community-level conventions may be inferred even with relatively short sparse observations from each partner. To resolve these questions, future work will need to develop new methods for eliciting children's expectations about partner-specificity and variability of meanings.

Third, our work suggests a new explanation for why young children may struggle to coordinate on *ad hoc* conventions with one another in repeated reference games (Glucksberg, Krauss, & Weisberg, 1966; Krauss & Glucksberg, 1969, 1977). Children as old as fifth grade only improve with assistance from the experimenter (Matthews, Lieven, & Tomasello, 2007): instead of beginning with the long indefinite descriptions that adults provide, it was observed that children began with shorter descriptions like *Mother's dress* (see also Kempe, Gauvrit, Gibson, & Jamieson, 2019). While these failures were initially attributed to limits on theory of mind use, this explanation has been complicated by findings that children cannot even interpret their own utterances after a delay (Asher & Oden, 1976). In other words, errors are not well-explained by egocentric adherence to a strongly preferred label, or downstream failures to acknowledge that this preferred label may not be understood by their partner; there appears to be no such preferred label.

Our model raises the possibility that the problem may instead stem from production with an impoverished lexical prior: there are no existing conventions for such a novel object in their vocabulary (Goldberg, 2019), leaving their preferences dispersed widely over "good-enough" constructions. Indeed, when children are paired with their caregivers rather than peers, they are able to successfully form conventions Leung, Hawkins, and Yurovsky (2020). Parents helped to interactively scaffold these conventions, both by proactively seeking clarification in the listener role (e.g. Anderson, Clark, & Mullin, 1994) and by providing more descriptive labels in the speaker role, which children adopted themselves on later trials. These results highlight one way in which our model of convention formation may coincide with models of language acquisition in development: in both cases, listeners are trying to infer the meanings of words in the speaker's lexicon (Bohn & Frank, 2019).

While we have highlighted three particular developmental phenomena where our approach may generate novel predictions, there are many finer-grained questions raised by our computational approach. For example, is the child's lexical expectations in communication best explained as a (resource-limited) representation of *others'* lexicons, or as an egocentric (asocial) epistemic state? To what extent is the ability to retrieve or use this lexical prior constrained by theory of mind development? Even infants are sensitive to coarse social distinctions based on foreign vs. native language (Kinzler, Dupoux, & Spelke, 2007), or accent (Kinzler, Shutts, DeJesus, & Spelke, 2009), but when do fully partner-specific representations develop?

The role of communication modality

While we have focused primarily on verbal and textual communication channels, there has beens significant progress understanding the dynamics of adaptation in other communication modalities, including graphical (Garrod et al., 2007; Theisen, Oberlander, & Kirby, 2010; Hawkins, Sano, et al., 2019) gestural (Fay, Lister, Ellison, & Goldin-Meadow, 2013; Motamedi, Schouwstra, Smith, Culbertson, & Kirby, 2019; Bohn, Kachel, & Tomasello, 2019) and other *de novo* modalities (Galantucci, 2005; Roberts & Galantucci, 2012; Roberts, Lewandowski, & Galantucci, 2015; Verhoef, Roberts, & Dingemanse, 2015; Verhoef, Walker, & Marghetis, 2016; Kempe et al., 2019). These modalities are important for our account in several ways.

Most importantly, it is a core claim of our hierarchical account that the basic learning mechanisms underlying adaptation and convention formation are domain-general. In other words, we predict that there is nothing inherently special about spoken or written language: any system that humans use to communicate should display similar *ad hoc* learning and convention formation dynamics because in every case people are simply trying to infer the system of meaning being used by their partner. Directly comparing behavior in repeated reference games across different modalities is therefore necessary to determine which adaptation effects, if any, are robust and attributable to modality-general mechanisms.

At the same time, our hierarchical learning model predicts a critical role for the *priors* we build up across interactions with many individuals. We therefore predict that different communication modalities should display certain systematic differences due to the representational structure of the communication channel. For example, in the verbal modality, the tangram shapes from Clark and Wilkes-Gibbs (1986) are highly "innominate" (Hupet, Seron, & Chantraine, 1991) – most people do not have much experience naming or describing them with words, so their priors are weak and local adaptation plays a greater role. In the graphical modality, where communication takes place by drawing on a shared sketchpad, people can be expected to have a stronger prior rooted in assumptions about shared perceptual systems and visual similarity (Fan, Yamins, & Turk-Browne, 2018) – drawing a quick sketch of the tangram's outline may suffice for understanding.

Other referents have precisely the opposite property: to distinguish between natural images of dogs, people may have strong existing conventions in the linguistic modality (e.g. 'husky', 'poodle', 'pug') but making the necessarily fine-grained visual distinctions in the graphical modality may be initially very costly for novices (Fan, Hawkins, Wu, & Goodman, 2020), requiring the formation of local conventions to achieve understanding (Hawkins, Sano, et al., 2019). The gestural modality also has its own distinctive prior, which also allows communicators to use time and the space around them to convey mimetic or depictive meanings that may be difficult to encode verbally or graphically (Goldin-Meadow & McNeill, 1999; Clark, 2016; McNeill, 1992). We suggest that differences in production and comprehension across modalities may therefore be understood by coupling modality-specific priors with modality-generic learning mechanisms.

The role of feedback and backchannels

If convention formation is grounded in inference, then an important corollary of our model is that the extent to which partners are able to coordinate should depend critically on the observations D_i they condition on in Eq. 4. In the *complete* absence of feedback — when the speaker is instructed to repeatedly refer to a set of objects for a listener who is not present and will do their half of the task offline — there is no reduction in message length Hupet and Chantraine (1992). Our simulations examined the most minimal sources of feedback: the speaker's utterance and the listener's response in a reference game. A key direction for future work is to

account for richer forms of feedback that arise in natural communication. For example, a key feature of dialogue is the capacity for a *real-time back-channel*. The listener may say anything at any point in time, thus allowing for interjections (uh-huh, hmmm, huh?), clarification questions, and other listener-initiated forms of feedback. Elaborating the generative model of the listener to include these verbal behaviors will be critical to explaining other key results from Clark and Wilkes-Gibbs (1986), including changes in the frequency of listener-initiated feedback over the course of interaction.

Additionally, such an elaboration would allow our model to address early empirical results from Krauss and Weinheimer (1966) manipulating the feedback channel: participants were able to talk bidirectionally in one condition but in another condition, the channel was unidirectional. The speaker was prevented from hearing the listener's responses. This feedback manipulation was crossed with a behavioral feedback manipulation where the experimenters intercepted the listener's responses: one group of speakers was told that their partner made the correct response 100% of the trials (regardless of their real responses), while another was told on half of the trials that their partner made the incorrect response. Our account of increasing efficiency (**P1**) predicts that speaker may not have sufficient evidence to justify shorter utterances in the absence of evidence about how their longer descriptions are being interpreted. Indeed, Krauss and Weinheimer (1966) found that both channel contribute independently to gains in efficiency. Speakers kept using longer utterances when they observed that their partner was making errors. But blocking the real-time verbal back-channel significantly limited reduction, even when told that their partner was achieving perfect accuracy. In the extreme case of trying to communicate to a listener who can't respond and also appears to not understand, speaker utterance length actually *increased* with repetition after an early dip.

More graded disruptions of feedback seem to force the speaker to use more words overall but not to significantly change the rate of reduction. For example, Krauss and Bricker (1967) tested a transmission delay to temporally shift feedback and an access delay to block the onset of listener feedback until the speaker is finished. Later, Krauss, Garlock, Bricker, and McMahon (1977) replicated the adverse effect of delay but showed that undelayed visual access to one's partner cancelled out the effect and returned the number of words used to baseline. On the listener's part, too, the ability to actively *give* feedback appears critical for coordination. Schober and Clark (1989) showed that even listeners who *overheard* the entire game were significantly less accurate than listeners who could directly interact with the speaker, even though they heard the exact same utterances, presumably because the speaker was not able to take their particular sources of confusion into account. Our model provide a formal framework to begin probing how these different forms of communicative feedback may license different inferences about one's partner in interaction.

The role of social knowledge

Real-world communities are much more complex than the simple networks we considered: each speaker takes part in a number of overlapping subcommunities. For example, we use partially distinct conventions depending on whether we are communicating with psychologists, friends from high school, bilinguals, or children (Auer, 2013). When a scientist is talking to other scientists about their work, they know they can use efficient technical shorthand that they would avoid when talking to their non-expert friends and family. Previous work has probed representations of community membership by manipulating the extent to which cultural background is shared between speaker and listener. For example, Isaacs and Clark (1987) paired participants who had either lived in NYC or had never been there for a task referring to landmarks in the city (e.g. "Rockefeller Center"). Within just a few utterances from a novel partner, people could infer whether they were playing with an expert or novice and immediately adjust their language use to be appropriate for this inferred identity. Social information about a partner's group can be so important that even players in artificial-language games react to the restrictions of social anonymity by learning to identify members of their community using distinctive signals (Roberts, 2010).

For future work using hierarchical Bayesian models to address the full scale of an individual's network of communities, additional social knowledge about these communities must be learned and represented in the generative model. Larger-scale networked experiments can be used to evaluate the hypothesis that a hierar-

chical representation of conventions includes not just a partner-specific level and population-wide level but also intermediate community levels. This hypothesis can be formalized by including additional latent representations of community membership into our hierarchical model. That is, in addition to updating our model of a particular *partner* based on immediate feedback, even sparse observations of a partner's language use may license much broader inferences about their lexicon via diagnostic information about their social group or background. If someone's favorite song is an obscure B-side from an 80s hardcore band, you can make fairly strong inferences about what else they like to listen to and how similar they might be to you (Vélez, Bridgers, & Gweon, 2019; Gershman, Pouncy, & Gweon, 2017). Similarly, if someone casually refers to an obscure New York landmark, you may be able to update your beliefs not just about that lexical item but about a number of other lexical conventions shared among New Yorkers. Lexica cluster within social groups, so inverting this relationship can yield rapid lexical learning from inferences about social group membership.

This explanation is also consistent with broader linguistic phenomena outside the realm of repeated reference games. For example, Potts and Levy (2015) showed that lexical uncertainty is critical for capturing constructions like *oenophile or wine lover*, where a disjunction of synonymous terms is taken to convey a definition – information about the speaker's lexicon – rather than a disjoint set. While the reasons that speakers produce such constructions are complex, we would expect that speakers will be more likely to produce the definitional *or* when the component word is expected to be rarer or more obscure for a particular partner: when there is additional uncertainty over its likely meaning in the listener's lexicon.

Process-level mechanisms for adaptation

Finally, while we have provided a computational-level account of coordination and convention formation in terms of hierarchical inference, there remain many possible process-level mechanisms that may perform this computation. In this section, we discuss two interlocking process-level questions: (1) exactly which representations are being adapted? (2) how does our model scale to larger spaces of utterances and referents?

Which representations are adapted? While our model formulation focused on adaptation at the level of the lexicon (i.e. inferences about ϕ, representing different possible lexical meanings), this is only one of many internal representations that may need to be adapted to achieve successful coordination. Three other possible representational bases have been explored in the literature.

First, it is possible that adaptation takes place upstream of the lexicon, directly implicating perceptual or *conceptual* representations (Garrod & Anderson, 1987; Healey, Swoboda, Umata, & King, 2007) That is, there may also be uncertainty about how a particular partner construes the referent itself, and communication may require constructing a shared, low-dimensional conceptual space where the relevant referents can be embedded (Stolk, Verhagen, & Toni, 2016). This is particularly clear in the classic maze task (Garrod & Anderson, 1987) where giving effective spatial directions requires speakers to coordinate on what spatial representations to use (e.g. paths, coordinates, lines, or landmarks).

Second, it is possible that adaptation takes place even further upstream, at the level of *social* representations (Jaech & Ostendorf, 2018). Rather than directly updating beliefs about lexical or conceptual representations, we may update a holistic representation of the partner themselves (e.g. as a "partner embedding" in a low-dimensional vector space) that is used to retrieve downstream conceptual and lexical representations. Under this representational scheme, the mapping from the social representation to particular conventions is static, and *ad hoc* adaptation is limited to learning where a particular partner belongs in the overall social space.

Third, it is possible that expectations about other lower-level features of a partner are also adapted through interaction. For example, interactive alignment accounts (Pickering & Garrod, 2004) have argued that activating phonetic or syntactic features that are associated with specific lemmas in the lexicon can percolate up to strengthen higher levels of representation (Roelofs, 1992; Pickering & Branigan, 1998). Thus, learning about a partner's word frequency (Louwerse, Dale, Bard, & Jeuniaux, 2012), syntax (Gruberg, Ostrand, Momma, & Ferreira, 2019; Levelt & Kelter, 1982), body postures (Lakin & Chartrand, 2003), speech rate (Giles, Coupland, & Coupland, 1991), or even informational complexity (Abney, Paxton, Dale, & Kello, 2014) could be functionally useful for commu-

nication if they covary with, or are informative about, higher-level representations.

Scalability to larger utterance and referent spaces. While a fully Bayesian formulation elegantly captures the inference problem at the core of our theory, the posterior update step in Eq. 4 grows increasingly intractable as the space of utterances and referents grows. This computational limitation is especially evident when considering how to quantitatively fit models to the natural-language data produced in unconstrained reference games, or applications in modern machine learning where we would like to build artificial agents that can adapt to human partners. Generalizing our framework to arbitrary natural language (e.g. referring expressions using the full vocabulary of an adult language user) and arbitrary visual input (e.g. sensory impressions of novel objects such as tangrams) not only requires a different representational space for language and visual input, it may require different inference algorithms.

The first computational obstacle is the lexical representation. A discrete $m \times n$ matrix containing entries for each utterance-referent pair, as typically used in convention formation simulations, has two primary limitations as a proxy for human representations: (1) it grows quadratically as additional utterances and referents are added while becoming increasingly sparse (i.e. most words only apply to a relatively small set of referents) and (2) it does not straightforwardly represent similarity relationships between utterances and referents (i.e. a new referent must be added as a distinct new column, even if it is visually similar to a known referent). This limitation is especially clear when referents are presented as raw images.

The second computational obstacle is the inference algorithm. Even if a more scalable family of lexical representations $\mathcal{L}_\phi$ is chosen, any parameterization ϕ of this representation will be much higher-dimensional than we have considered. For example, if (Θ, ϕ) are defined to be the weights of a neural network, then maintaining uncertainty $P(\Theta, \phi)$ would require placing a prior over the weights and the posterior update $P(\Theta, \phi|D)$ would require a difficult integral over a very high-dimensional support. This is called a (hierarchical) Bayesian neural network (MacKay, 1992; Neal, 1996). While recent algorithmic breakthroughs have made (approximate) inference for such networks increasingly tractable (e.g. Hernández-Lobato & Adams, 2015; Lacoste et al., 2018; Dusenberry et al., 2020; Izmailov et al., 2020), they remain untested as cognitive models.

While these obstacles are significant, we argue that our hierarchical Bayesian framework is nonetheless well-poised to explain coordination and convention-formation at scale. In particular, recent formal connections between hierarchical Bayes and gradient-based meta-learning approaches in machine learning (Grant, Finn, Levine, Darrell, & Griffiths, 2018) suggest an alternative algorithm that relaxes the full prior over Θ to a point estimate and replaces the difficult integral in the posterior update with a handful of (regularized) gradient steps. This more tractable algorithm approximates the same computational-level problem we have formulated but provides a different perspective: conventions are learned *initializations* and coordination is partner-specific *fine-tuning* or *domain adaptation* of vector representations. The fine-tuning approach, initializing with a state-of-the-art neural language model, has recently accounted for psycholinguistic data on reading times (Van Schijndel & Linzen, 2018) as well as *ad hoc* convention formation in reference games using (unseen) natural images as referents (Hawkins, Kwon, et al., 2020).

Conclusion. In summary, we have argued that language is not a rigid body of knowledge that we acquire at an early age and deploy mechanically for the rest of our lives. Nor is language change entirely explained by slow, inter-generational drift. Language is a means for communication – a shared interface between minds – and must therefore adapt over the rapid timescales required by communication. As new *ad hoc* concepts arise, new *ad hoc* conventions must be formed. In other words, we are constantly learning language. Not just one language, but a family of related languages, across every repeated interaction with every partner.

> Let us conclude not that 'there is no such thing as a language' that we bring to interaction with others. Say rather that there is no such thing as the one total language that we bring. We bring numerous only loosely connected languages from the loosely connected communities that we inhabit. (Hacking, 1986)

References

Abney, D. H., Paxton, A., Dale, R., & Kello, C. T. (2014). Complexity matching in dyadic conversation. *Journal of Experimental Psychology: General*, *143*(6), 2304.

Anderson, A. H., Clark, A., & Mullin, J. (1994). Interactive communication between children: learning how to make language work in dialogue. *Journal of Child Language*, *21*(2), 439–463.

Anderson, J. R., & Schooler, L. J. (2000). *The adaptive nature of memory.* Oxford University Press.

Angela, J. Y., & Cohen, J. D. (2009). Sequential effects: superstition or rational behavior? In *Advances in Neural Information Processing Systems* (pp. 1873–1880).

Asher, S. R., & Oden, S. L. (1976). Children's failure to communicate: An assessment of comparison and egocentrism explanations. *Developmental Psychology*, *12*(2), 132.

Auer, P. (2013). *Code-switching in conversation: Language, interaction and identity*. Routledge.

Baronchelli, A., Gong, T., Puglisi, A., & Loreto, V. (2010). Modeling the emergence of universality in color naming patterns. *Proceedings of the National Academy of Sciences*, *107*(6), 2403–2407.

Baronchelli, A., Loreto, V., & Steels, L. (2008). In-depth analysis of the naming game dynamics: the homogeneous mixing case. *International Journal of Modern Physics C*, *19*(05), 785–812.

Barr, D. J. (2004). Establishing conventional communication systems: Is common knowledge necessary? *Cognitive Science*, *28*(6), 937–962.

Bergen, L., Levy, R., & Goodman, N. (2016). Pragmatic reasoning through semantic inference. *Semantics and Pragmatics*, *9*(20).

Berniker, M., & Kording, K. (2008). Estimating the sources of motor errors for adaptation and generalization. *Nature Neuroscience*, *11*(12), 1454.

Beuls, K., & Steels, L. (2013). Agent-based models of strategies for the emergence and evolution of grammatical agreement. *PloS one*, *8*(3), e58960.

Bicchieri, C. (2006). *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press.

Bloom, P. (2002). *How children learn the meanings of words*. MIT press.

Bohn, M., & Frank, M. C. (2019). The pervasive role of pragmatics in early language. *Annual Review of Developmental Psychology*, *1*, 223–249.

Bohn, M., Kachel, G., & Tomasello, M. (2019). Young children spontaneously recreate core properties of language in a new modality. *Proceedings of the National Academy of Sciences*, *116*(51), 26072–26077.

Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(6), 1482.

Brennan, S. E., & Hanna, J. E. (2009). Partner-specific adaptation in dialog. *Topics in Cognitive Science*, *1*(2).

Brown-Schmidt, S. (2009). Partner-specific interpretation of maintained referential precedents during interactive dialog. *Journal of Memory and Language*, *61*(2), 171–190.

Bruner, J., Goodnow, J., & Austin, G. (1956). *A study of thinking*. New York: John Wiley & Sons.

Bruner, J., O'Connor, C., Rubin, H., & Huttegger, S. M. (2014). David Lewis in the lab: Experimental results on the emergence of meaning. *Synthese*, 1–19.

Carroll, J. M. (1980). Naming and describing in social communication. *Language and Speech*, *23*(4), 309–322.

Centola, D., & Baronchelli, A. (2015). The spontaneous emergence of conventions: An experimental study of cultural evolution. *Proceedings of the National Academy of Sciences*, *112*(7), 1989–1994.

Clark, E. V. (2009). *First language acquisition*. Cambridge University Press.

Clark, H. H. (1996). *Using language*. Cambridge university press Cambridge.

Clark, H. H. (1998). Communal lexicons. In K. Malmkjaer & J. Williams (Eds.), *Context in language learning and language understanding* (pp. 63–87). Cambridge: Cambridge University Press.

Clark, H. H. (2016). Depicting as a method of communication. *Psychological review*, *123*(3), 324.

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, *22*(1), 1–39.

Clopper, C. G., & Pisoni, D. B. (2004). Effects of talker variability on perceptual learning of di-

alects. *Language and Speech*, *47*(3), 207–238.

Collins, A. M., & Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*, *8*(2), 240–247.

Cowans, P. J. (2004). Information retrieval using hierarchical dirichlet processes. In *Proceedings of the 27th conference on research and development in information retrieval* (pp. 564–565).

Creel, S. C., & Bregman, M. R. (2011). How talker identity relates to language processing. *Language and Linguistics Compass*, *5*(5), 190–204.

Dale, R., & Reiter, E. (1995). Computational interpretations of the gricean maxims in the generation of referring expressions. *Cognitive Science*, *19*(2), 233–263.

Dautriche, I., Goupil, L., Smith, K., & Rabagliati, H. (2021). Knowing how you know: Toddlers reevaluate words learned from an unreliable speaker. *Open Mind*, *5*, 1–19.

Davidson, D. (1984). Communication and convention. *Synthese*, 3–17.

Davidson, D. (1986). A nice derangement of epitaphs. *Philosophical grounds of rationality: Intentions, categories, ends*, *4*, 157–174.

Davidson, D. (1994). The social aspect of language. In B. McGuiness & G. Oliveri (Eds.), *The philosophy of Michael Dummett* (pp. 1–16).

Degen, J., Hawkins, R. D., Graf, C., Kreiss, E., & Goodman, N. D. (2020). When redundancy is useful: A bayesian approach to "overinformative" referring expressions. *Psychological Review*, *127*(4), 591–621.

Dummett, M. (1994). Reply to Davidson. In B. McGuiness & G. Oliveri (Eds.), *The philosophy of Michael Dummett* (p. 257-67).

Dusenberry, M., Jerfel, G., Wen, Y., Ma, Y., Snoek, J., Heller, K., ... Tran, D. (2020). Efficient and scalable bayesian neural nets with rank-1 factors. In *International conference on machine learning* (pp. 2782–2792).

Eckert, P. (2012). Three waves of variation study: The emergence of meaning in the study of sociolinguistic variation. *Annual review of Anthropology*, *41*, 87–100.

Erev, I., & Roth, A. E. (1998). Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria. *American economic review*, 848–881.

Fan, J. E., Hawkins, R. D., Wu, M., & Goodman, N. D. (2020). Pragmatic inference and visual abstraction enable contextual flexibility during visual communication. *Computational Brain & Behavior*, *3*(1), 86–101.

Fan, J. E., Yamins, D. L., & Turk-Browne, N. B. (2018). Common object representations for visual production and recognition. *Cognitive Science*, *42*(8), 2670–2698.

Fay, N., Garrod, S., Roberts, L., & Swoboda, N. (2010). The interactive evolution of human communication systems. *Cognitive Science*, *34*(3).

Fay, N., Lister, C., Ellison, T., & Goldin-Meadow, S. (2013). Creating a communication system from scratch: gesture beats vocalization hands down. *Frontiers in Psychology*, *5*, 354–354.

Frank, M. C. (2017). What's the relationship between language and thought? the optimal semantic expressivity hypothesis. Retrieved from `http://babieslearninglanguage.blogspot.com/2017/07/whats-relationship-between-language-and.html`

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998–998.

Frank, M. C., Goodman, N. D., & Tenenbaum, J. B. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, *20*(5), 578–585.

Franke, M., & Jäger, G. (2012). Bidirectional optimization from reasoning and learning in games. *Journal of Logic, Language and Information*, *21*(1), 117–139.

Franke, M., & Jäger, G. (2016). Probabilistic pragmatics, or why bayes' rule is probably important for pragmatics. *Zeitschrift für sprachwissenschaft*, *35*(1), 3–44.

Fraser, B. (2010). Pragmatic competence: The case of hedging. *New approaches to hedging*, *1534*.

Fudenberg, D., & Levine, D. K. (2014). Recency, consistent learning, and Nash equilibrium. *Proceedings of the National Academy of Sciences*, *111*(Supplement 3), 10826–10829.

Fussell, S. R., & Krauss, R. M. (1989). The effects of intended audience on message production and comprehension: Reference in a common ground

framework. *Journal of Experimental Social Psychology*, *25*(3), 203–219.

Galantucci, B. (2005). An experimental study of the emergence of human communication systems. *Cognitive Science*, *29*(5), 737–767.

Garrod, S., & Anderson, A. (1987). Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition*, *27*(2), 181–218.

Garrod, S., & Doherty, G. (1994). Conversation, coordination and convention: An empirical investigation of how groups establish linguistic conventions. *Cognition*, *53*(3).

Garrod, S., Fay, N., Lee, J., Oberlander, J., & MacLeod, T. (2007). Foundations of representation: where might graphical symbol systems come from? *Cognitive Science*, *31*(6), 961–987.

Garrod, S., & Pickering, M. J. (2009). Joint action, interactive alignment, and dialog. *Topics in Cognitive Science*, *1*(2), 292–304.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (3rd ed.). CRC press Boca Raton, FL.

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Gershman, S. J. (2017). On the blessing of abstraction. *The Quarterly Journal of Experimental Psychology*, *70*(3), 361–365.

Gershman, S. J., Pouncy, H. T., & Gweon, H. (2017). Learning the structure of social influence. *Cognitive Science*, *41*, 545–575.

Gibson, E., Futrell, R., Jara-Ettinger, J., Mahowald, K., Bergen, L., Ratnasingam, S., ... Conway, B. R. (2017). Color naming across languages reflects color use. *Proceedings of the National Academy of Sciences*, *114*(40), 10785–10790.

Gibson, E., Futrell, R., Piandadosi, S. T., Dautriche, I., Mahowald, K., Bergen, L., & Levy, R. (2019). How efficiency shapes human language. *Trends in Cognitive Sciences*, *23*(5), 389–407.

Giles, H., Coupland, N., & Coupland, J. (1991). *Contexts of accommodation: Developments in applied sociolinguistics*. Cambridge University Press.

Glucksberg, S., Krauss, R. M., & Weisberg, R. (1966). Referential communication in nursery school children: Method and some preliminary findings. *Journal of Experimental Child Psychology*, *3*(4), 333–342.

Goldberg, A. E. (2019). *Explain me this: Creativity, competition, and the partial productivity of constructions*. Princeton University Press.

Goldin-Meadow, S., & McNeill, D. (1999). The role of gesture and mimetic representation in making language the province of speech. *The descent of mind: Psychological perspectives on hominid evolution.*, 155–172.

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

Goodman, N. D., & Stuhlmüller, A. (electronic). *The design and implementation of probabilistic programming languages.* Retrieved 2015/1/16, from `http://dippl.org`

Goodman, N. D., Ullman, T. D., & Tenenbaum, J. B. (2011). Learning a theory of causality. *Psychological Review*, *118*(1), 110.

Graesser, L., Cho, K., & Kiela, D. (2019). Emergent linguistic phenomena in multi-agent communication games. *arXiv preprint arXiv:1901.08706*.

Grant, E., Finn, C., Levine, S., Darrell, T., & Griffiths, T. (2018). Recasting gradient-based meta-learning as Hierarchical Bayes. In *6th International Conference on Learning Representations.*

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Syntax and semantics* (pp. 43–58). New York: Academic Press.

Gruberg, N., Ostrand, R., Momma, S., & Ferreira, V. S. (2019). Syntactic entrainment: The repetition of syntactic structures in event descriptions. *Journal of Memory and Language*, *107*, 216–232.

Gulordava, K., Brochhagen, T., & Boleda, G. (2020). Which one is the dax? achieving mutual exclusivity with neural networks. *arXiv preprint arXiv:2004.03902*.

Hacking, I. (1986). The parody of conversation. In E. LePore (Ed.), *Truth and interpretation: Perspectives on the philosophy of donald davidson* (p. 447-458). Cambridge.

Hawkins, R. D. (2015). Conducting real-time multiplayer experiments on the web. *Behavior Research Methods*, *47*(4), 966-976.

Hawkins, R. D., Frank, M. C., & Goodman, N. D. (2017). Convention-formation in iterated reference games. In *Proceedings of the 39th annual meeting of the cognitive science society.*

Hawkins, R. D., Frank, M. C., & Goodman, N. D. (2020). Characterizing the dynamics of learning in repeated reference games. *Cognitive Science*, *44*(6), e12845.

Hawkins, R. D., Franke, M., Smith, K., & Goodman, N. D. (2018). Emerging abstractions: Lexical conventions are shaped by communicative context. In *Proceedings of the 40th annual meeting of the cognitive science society.*

Hawkins, R. D., & Goldstone, R. L. (2016, 03). The formation of social conventions in real-time environments. *PLoS ONE*, *11*(3), 1-14.

Hawkins, R. D., Goodman, N. D., Goldberg, A. E., & Griffiths, T. L. (2020). Generalizing meanings from partners to populations: Hierarchical inference supports convention formation on networks. In *Proceedings of the 42th annual meeting of the cognitive science society.*

Hawkins, R. D., Goodman, N. D., & Goldstone, R. L. (2019). The emergence of social norms and conventions. *Trends in Cognitive Sciences*, *23*(2), 158–169.

Hawkins, R. D., Kwon, M., Sadigh, D., & Goodman, N. D. (2020). Continual adaptation for efficient machine communication. *Proceedings of the 24th Conference on Computational Natural Language Learning*.

Hawkins, R. D., Sano, M., Goodman, N. D., & Fan, J. E. (2019). Disentangling contributions of visual information and interaction history in the formation of graphical conventions. In *Proceedings of the 41st annual meeting of the cognitive science society* (pp. 415–421).

Healey, P. G., Swoboda, N., Umata, I., & King, J. (2007). Graphical language games: Interactional constraints on representational form. *Cognitive Science*, *31*(2), 285–309.

Heck, R. K. (2006). Idiolects. In J. J. Thomson & A. Byrne (Eds.), *Content and modality: Themes from the philosophy of Robert Stalnaker* (p. 61-92).

Hernández-Lobato, J. M., & Adams, R. (2015). Probabilistic backpropagation for scalable learning of bayesian neural networks. In *International conference on machine learning* (pp. 1861–1869).

Hupet, M., & Chantraine, Y. (1992). Changes in repeated references: Collaboration or repetition effects? *Journal of Psycholinguistic Research*, *21*(6), 485–496.

Hupet, M., Seron, X., & Chantraine, Y. (1991). The effects of the codability and discriminability of the referents on the collaborative referring procedure. *British Journal of Psychology*, *82*(4), 449–462.

Hurford, J. R. (1989). Biological evolution of the saussurean sign as a component of the language acquisition device. *Lingua*, *77*(2), 187–222.

Isaacs, E. A., & Clark, H. H. (1987). References in conversation between experts and novices. *Journal of Experimental Psychology: General*, *116*(1), 26.

Izmailov, P., Maddox, W. J., Kirichenko, P., Garipov, T., Vetrov, D., & Wilson, A. G. (2020). Subspace inference for bayesian deep learning. In *Uncertainty in artificial intelligence* (pp. 1169–1179).

Jaech, A., & Ostendorf, M. (2018). Low-rank rnn adaptation for context-aware language modeling. *Transactions of the Association for Computational Linguistics*, *6*, 497–510.

Jäger, G. (2007). The evolution of convex categories. *Linguistics and Philosophy*, *30*(5), 551–564.

Jäger, G., & Van Rooij, R. (2007). Language structure: Psychological and social constraints. *Synthese*, *159*(1), 99–130.

Kalm, K., & Norris, D. (2018). Visual recency bias is explained by a mixture model of internal representations. *Journal of Vision*, *18*(7), 1–1.

Kao, J. T., Wu, J. Y., Bergen, L., & Goodman, N. D. (2014). Nonliteral understanding of number words. *Proceedings of the National Academy of Sciences*, *111*(33), 12002–12007.

Kemp, C., Goodman, N. D., & Tenenbaum, J. B. (2010). Learning to learn causal models. *Cognitive Science*, *34*(7), 1185–1243.

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental Science*, *10*(3), 307–321.

Kemp, C., & Regier, T. (2012). Kinship categories across languages reflect general communicative

principles. *Science*, *336*(6084), 1049–1054.

Kemp, C., Xu, Y., & Regier, T. (2018). Semantic typology and efficient communication. *Annual Review of Linguistics*, *4*, 109–128.

Kempe, V., Gauvrit, N., Gibson, A., & Jamieson, M. (2019). Adults are more efficient in creating and transmitting novel signalling systems than children. *Journal of Language Evolution*, *4*(1), 44–70.

Kinzler, K. D., Dupoux, E., & Spelke, E. S. (2007). The native language of social cognition. *Proceedings of the National Academy of Sciences*, *104*(30), 12577–12580.

Kinzler, K. D., Shutts, K., DeJesus, J., & Spelke, E. S. (2009). Accent trumps race in guiding children's social preferences. *Social Cognition*, *27*(4), 623.

Kirby, S., Tamariz, M., Cornish, H., & Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. *Cognition*, *141*, 87–102.

Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review*, *122*(2), 148.

Koenig, M. A., & Woodward, A. L. (2010). Sensitivity of 24-month-olds to the prior inaccuracy of the source: possible mechanisms. *Developmental Psychology*, *46*(4), 815.

Krauss, R. M., & Bricker, P. D. (1967). Effects of transmission delay and access delay on the efficiency of verbal communication. *The Journal of the Acoustical Society of America*, *41*(2), 286–292.

Krauss, R. M., Garlock, C. M., Bricker, P. D., & McMahon, L. E. (1977). The role of audible and visible back-channel responses in interpersonal communication. *Journal of Personality and Social Psychology*, *35*(7), 523.

Krauss, R. M., & Glucksberg, S. (1969). The development of communication: Competence as a function of age. *Child Development*, 255–266.

Krauss, R. M., & Glucksberg, S. (1977). Social and nonsocial speech. *Scientific American*, *236*(2), 100–105.

Krauss, R. M., & Weinheimer, S. (1964). Changes in reference phrases as a function of frequency of usage in social interaction: A preliminary study. *Psychonomic Science*, *1*(1-12), 113–114.

Krauss, R. M., & Weinheimer, S. (1966). Concurrent feedback, confirmation, and the encoding of referents in verbal communication. *Journal of Personality and Social Psychology*, *4*(3), 343.

Lacoste, A., Oreshkin, B., Chung, W., Boquet, T., Rostamzadeh, N., & Krueger, D. (2018). Uncertainty in multitask transfer learning. *arXiv preprint arXiv:1806.07528*.

Lakin, J. L., & Chartrand, T. L. (2003). Using nonconscious behavioral mimicry to create affiliation and rapport. *Psychological Science*, *14*(4), 334–339.

Lassiter, D., & Goodman, N. D. (2015). Adjectival vagueness in a bayesian model of interpretation. *Synthese*, 1-36.

Lazaridou, A., Kuncoro, A., Gribovskaya, E., Agrawal, D., Liska, A., Terzi, T., ... others (2021). Pitfalls of static language modelling. *arXiv preprint arXiv:2102.01951*.

Lepore, E., & Ludwig, K. (2007). The reality of language: On the davidson/dummett exchange. *The philosophy of Michael Dummett*, 185–214.

Leung, A., Hawkins, R. D., & Yurovsky, D. (2020). Parents scaffold the formation of conversational pacts with their children. In *Proceedings of the 42nd annual conference of the cognitive science society.*

Lev-Ari, S. (2017). Talking to fewer people leads to having more malleable linguistic representations. *PloS One*, *12*(8), e0183593.

Levelt, W. J., & Kelter, S. (1982). Surface form and memory in question answering. *Cognitive psychology*, *14*(1), 78–106.

Lewis, D. (1969). *Convention: A philosophical study*. Harvard University Press.

Louwerse, M. M., Dale, R., Bard, E. G., & Jeuniaux, P. (2012). Behavior matching in multimodal communication is synchronized. *Cognitive Science*, *36*(8), 1404–1426.

Luchkina, E., Sobel, D. M., & Morgan, J. L. (2018). Eighteen-month-olds selectively generalize words from accurate speakers to novel contexts. *Developmental Science*, *21*(6), e12663.

Luke, S. G. (2017). Evaluating significance in linear mixed-effects models in R. *Behavior research methods*, *49*(4), 1494–1502.

MacKay, D. J. (1992). A practical bayesian framework for backpropagation networks. *Neural Computation*, *4*(3), 448–472.

Matthews, D., Lieven, E., & Tomasello, M. (2007). How toddlers and preschoolers learn to uniquely identify referents for others: A training study. *Child Development*, *78*(6), 1744–1759.

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. University of Chicago press.

Medlock, B., & Briscoe, T. (2007). Weakly supervised learning for hedge classification in scientific literature. In *ACL* (pp. 992–999).

Metzing, C., & Brennan, S. E. (2003). When conceptual pacts are broken: Partner-specific effects on the comprehension of referring expressions. *Journal of Memory and Language*, *49*(2).

Mordatch, I., & Abbeel, P. (2017). Emergence of grounded compositional language in multi-agent populations. *arXiv preprint arXiv:1703.04908*.

Moreno, M., & Baggio, G. (2014). Role asymmetry and code transmission in signaling games: An experimental and computational investigation. *Cognitive Science*, *39*(5), 918–943.

Motamedi, Y., Schouwstra, M., Smith, K., Culbertson, J., & Kirby, S. (2019). Evolving artificial sign languages in the lab: From improvised gesture to systematic sign. *Cognition*, *192*, 103964.

Neal, R. M. (1996). *Bayesian learning for neural networks* (Vol. 118). Springer Science & Business Media.

Ohmer, X., König, P., & Franke, M. (2020). Reinforcement of semantic representations in pragmatic agents leads to the emergence of a mutual exclusivity bias. *Proceedings of the 42nd Annual Cognitive Science Conference*.

Pearl, L., Goldwater, S., & Steyvers, M. (2010). Online learning mechanisms for bayesian models of word segmentation. *Research on Language and Computation*, *8*(2-3), 107–132.

Pickering, M. J., & Branigan, H. P. (1998). The representation of verbs: Evidence from syntactic priming in language production. *Journal of Memory and Language*, *39*(4), 633–651.

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and brain sciences*, *27*(2), 169–190.

Pickering, M. J., & Garrod, S. (2006). Alignment as the basis for successful communication. *Research on Language and Computation*, *4*(2-3), 203–228.

Polyn, S. M., Norman, K. A., & Kahana, M. J. (2009). A context maintenance and retrieval model of organizational processes in free recall. *Psychological Review*, *116*(1), 129.

Potts, C. (2019). A case for deep learning in semantics: Response to Pater. *Language*, *95*(1), e115–e124.

Potts, C., & Levy, R. (2015). Negotiating lexical uncertainty and speaker expertise with disjunction. In *Proceedings of the 41st annual meeting of the Berkeley Linguistics Society* (Vol. 41).

Qing, C., & Franke, M. (2015). Variations on a bayesian theme: Comparing bayesian models of referential reasoning. In *Bayesian natural language semantics and pragmatics* (pp. 201–220). Springer.

Raviv, L., Meyer, A., & Lev-Ari, S. (2019). Larger communities create more systematic languages. *Proceedings of the Royal Society B*, *286*(1907), 20191262.

Regier, T. (2005). The emergence of words: Attentional learning in form and meaning. *Cognitive Science*, *29*(6), 819–865.

Regier, T., Kemp, C., & Kay, P. (2015). 11 word meanings across languages support efficient communication. *The handbook of language emergence*, *87*, 237.

Roberts, G. (2010). An experimental study of social selection and frequency of interaction in linguistic diversity. *Interaction Studies*, *11*(1), 138–159.

Roberts, G., & Galantucci, B. (2012). The emergence of duality of patterning: Insights from the laboratory. *Language and Cognition*, *4*(4), 297–318.

Roberts, G., Lewandowski, J., & Galantucci, B. (2015). How communication changes when we cannot mime the world: Experimental evidence for the effect of iconicity on combinatoriality. *Cognition*, *141*, 52–66.

Roelofs, A. (1992). A spreading-activation theory of lemma retrieval in speaking. *Cognition*, *42*(1-3), 107–142.

Schober, M. F., & Clark, H. H. (1989). Understanding by addressees and overhearers. *Cognitive Psychology*, *21*(2), 211–232.

Scontras, G., Tessler, M., & Franke, M. (2018). Probabilistic language understanding: An in-

troduction to the rational speech act framework. Retrieved 2021-2-15, from https://www.problang.org

Shannon, C. E. (1948). A mathematical theory of communication. *Bell system technical journal*, *27*.

Siskind, J. M. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, *61*(1-2), 39–91.

Skyrms, B. (2010). *Signals: Evolution, learning, and information*. Oxford University Press.

Smith, K., Perfors, A., Fehér, O., Samara, A., Swoboda, K., & Wonnacott, E. (2017). Language learning, language use and the evolution of linguistic variation. *Philosophical Transactions of the Royal Society B*, *372*(1711), 20160051.

Smith, L. B., Suanda, S. H., & Yu, C. (2014). The unrealized promise of infant statistical word–referent learning. *Trends in Cognitive Sciences*, *18*(5), 251–258.

Smith, N. J., Goodman, N., & Frank, M. (2013). Learning and using language via recursive pragmatic reasoning about other agents. In *Advances in neural information processing systems* (pp. 3039–3047).

Sperber, D., & Wilson, D. (1986). *Relevance: Communication and cognition*. Harvard University Press.

Spike, M., Stadler, K., Kirby, S., & Smith, K. (2017). Minimal requirements for the emergence of learned signaling. *Cognitive Science*, *41*(3), 623–658.

Steels, L. (1995). A self-organizing spatial vocabulary. *Artificial Life*, *2*(3), 319–332.

Steels, L. (2011). Modeling the cultural evolution of language. *Physics of Life Reviews*, *8*(4), 339–356.

Steels, L. (2012). *Experiments in cultural language evolution* (Vol. 3). John Benjamins Publishing.

Steels, L. (2016). Agent-based models for the emergence and evolution of grammar. *Philosophical Transactions of the Royal Society B*, *371*(1701), 20150447.

Steels, L., & Belpaeme, T. (2005). Coordinating perceptually grounded categories through language: A case study for colour. *Behavioral and Brain Sciences*, *28*(4), 469–488.

Stolk, A., Verhagen, L., & Toni, I. (2016). Conceptual alignment: How brains achieve mutual understanding. *Trends in Cognitive Sciences*, *20*(3), 180–191.

Strawson, P. F. (1950). On referring. *Mind*, *59*(235), 320–344.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *science*, *331*(6022), 1279–1285.

Tessler, M. H., & Goodman, N. D. (2018). The language of generalization. *Psychological Review*.

Theisen, C. A., Oberlander, J., & Kirby, S. (2010). Systematicity and arbitrariness in novel communication systems. *Interaction Studies*, *11*(1), 14–32.

Tieleman, O., Lazaridou, A., Mourad, S., Blundell, C., & Precup, D. (2019). Shaping representations through communication: community size effect in artificial learning systems. *arXiv preprint arXiv:1912.06208*.

van Deemter, K. (2016). *Computational models of referring: A study in cognitive science*. MIT Press.

Van Fraassen, B. C. (1966). Singular terms, truth-value gaps, and free logic. *The Journal of Philosophy*, *63*(17), 481–495.

Van Schijndel, M., & Linzen, T. (2018). A neural model of adaptation in reading. *arXiv preprint arXiv:1808.09930*.

Vélez, N., Bridgers, S., & Gweon, H. (2019). The rare preference effect: Statistical information influences social affiliation judgments. *Cognition*, *192*, 103994.

Verhoef, T., Roberts, S. G., & Dingemanse, M. (2015). Emergence of systematic iconicity: Transmission, interaction and analogy. In D. C. Noelle et al. (Eds.), *Proceedings of the 37th annual conference of the cognitive science society.*

Verhoef, T., Walker, E., & Marghetis, T. (2016). Cognitive biases and social coordination in the emergence of temporal language. In *The 38th annual meeting of the cognitive science society (cogsci 2016)* (pp. 2615–2620).

Weber, R. A., & Camerer, C. F. (2003). Cultural conflict and merger failure: An experimental approach. *Management Science*, *49*(4), 400–415.

Wilkes-Gibbs, D., & Clark, H. H. (1992). Coordinating beliefs in conversation. *Journal of Memory and Language*, *31*(2), 183–194.

Winters, J., Kirby, S., & Smith, K. (2014). Languages adapt to their contextual niche. *Language and Cognition*, 1–35.

Winters, J., Kirby, S., & Smith, K. (2018). Contextual predictability shapes signal autonomy. *Cognition*, *176*, 15–30.

Wittgenstein, L. (1953). *Philosophical investigations*. Macmillan Publishing Company.

Wixted, J. T., & Ebbesen, E. B. (1991). On the form of forgetting. *Psychological Science*, *2*(6), 409–415.

Xu, F., & Tenenbaum, J. B. (2007). Word learning as bayesian inference. *Psychological Review*, *114*(2), 245.

Young, H. P. (1996). The economics of convention. *The Journal of Economic Perspectives*, *10*(2), 105–122.

Young, H. P. (2015). The Evolution of Social Norms. *Annual Review of Economics*, *7*, 359–387.

Yurovsky, D., & Frank, M. C. (2015). An integrative account of constraints on cross-situational learning. *Cognition*, *145*, 53–62.

Zaslavsky, N., Hu, J., & Levy, R. P. (2020). A rate-distortion view of human pragmatic reasoning. *arXiv preprint arXiv:2005.06641*.

Appendix A: Details of RSA model

Our setting poses several technical challenges for the Rational Speech Act (RSA) framework. In this Appendix, we describe these challenges in more detail and justify our choices.

Handling degenerate lexicons

First, when we allow the full space of possible lexicons ϕ, we must confront degenerate lexicons where an utterance u is literally false of every object in context, i.e. where $\mathcal{L}_\phi(o, u) = 0$ for all $o \in C$. In this case, the normalizing constant in Eq. 2 is zero, and the literal listener distribution is not well-defined. A similar problem may arise when no utterance in the speaker's repertoire is true of the target, in which case the S_1 distribution is not well-defined.

Several solutions to this problem were outlined by Bergen et al. (2016). One of these solutions is to use a 'softer' semantics in the literal listener, where a Boolean value of false does not strictly rule out an object but instead assigns a very low numerical score, e.g.

$$\mathcal{L}_\phi(o, u) = \begin{cases} 1 & \text{if } o \in \phi(u) \\ \epsilon & \text{o.w.} \end{cases}$$

Whenever there is at least one $o \in C$ where u is true, this formulation assigns negligible listener probability to objects where u is false, but ensures that the normalization constant to is non-zero even when u is false for all objects.

While this solution suffices for one-shot pragmatics under lexical uncertainty, where ϵ may be calibrated to be appropriately large, it runs into several technical complications in an iterated setting. First, due to numerical overflow at sufficiently high values of w_L and w_S at later iterations, elements may drop entirely out of the support at higher levels of recursion (e.g. L_1), leading the normalization constant to return to zero. Second, this 'soft' semantics creates unexpected and unintuitive consequences at the level of the pragmatic speaker. After renormalization in L_0, an utterance u that fails to refer to any object in context is also by definition equally *successful* for all objects (i.e. evaluating to ϵ for every object), leading to a uniform selection distribution. Consequently, S_1 may in some cases prefer utterances that are literally false of the target just as much as utterances which are true.

Instead of injecting ϵ into the lexical meaning, we ensure that the normalization constant is well-defined by adapting another method suggested by Bergen et al. (2016). First, we add a 'null' object to every context so that, even if a particular utterance is false of every real object in context, it will still apply to the null object, assigning the true target a negligible probability of being chosen. Intuitively, this null object can be interpreted as recognizing that the referring expression has a referent but it is not in context. Second, we add an explicit noise model at every level of recursion. That is, we assume every agent has a probability ϵ of choosing a random element of their support, ensuring a fixed non-zero floor on the likelihood of each element that is constant across levels of recursion. Formally this corresponds to a mixture distribution, e.g.

$$L_0^\epsilon(o|u, \phi) = \epsilon \cdot P_{unif}(o) + (1 - \epsilon) \cdot L_0(o|u, \phi)$$

$$S_1^\epsilon(u|o, \phi) = \epsilon \cdot P_{unif}(u) + (1 - \epsilon) \cdot S_1(u|o, \phi)$$

	Parameter	Example parameter settings
Partner design	What feedback is provided?	- no feedback at all - only correct/incorrect - real-time responses from partner
	Are you playing with the same partner?	- same partner for whole game - swap out partners every round - swap after k rounds
	What do you know about your partner?	- anonymous stranger - stranger with perceptual information - close friend
	How consistent are roles across repetitions?	- consistent director/matcher - alternate roles each round
Stimulus design	How familiar are targets?	- very familiar: colors, household objects - not at all familiar: tangrams, novel line drawings
	How complex are targets?	- very complex: busy visual scenes, clips of music - not at all complex: geometric drawings
	How consistent are targets across repetitions?	- exact same image of object - different pose/view of same object - different objects from same neighborhood
Context design	How similar are distractors to the target?	- very similar: same basic-level category - not at all similar: other categories
	What is the size of context?	- between 2 and 21
	How consistent is context across repetitions?	- exact same context each round - randomized context (sometimes far, sometimes close)
Repetition design	How many repetitions per target?	- between 3 and 100
	What is spacing between repetitions?	- block structure - sequential structure with interspersed contexts
Modality design	What medium is used for communication?	- text - audio - gesture - drawing

Table A1
Proposed parameterization for repeated reference games, each of which theoretically impacts the formation of conventions.

Marginalizing over ϕ_k

Another theoretical question arises about exactly how speaker and listener agents ought to marginalize over their uncertainty about ϕ_k when selecting actions (Eq. 3). In our formulation, the expectation is naturally taken over the entire *utility* each agent is using to act, i.e. if the speaker and listener utilities are defined to be

$$\begin{aligned} U_L(o;u,\phi_k) &= \log S_1(u|o,\phi_k) \\ U_S(u;o,\phi_k) &= (1-w_C)\log L_0(o|u,\phi_k) - w_C \cdot c(u) \end{aligned}$$

then the expectation is taken as follow:

$$\begin{aligned} L(o|u) &\propto \exp\left\{w_L \int P_L(\phi_k|D_k)\cdot U_L(u;o,\phi_k)\,d\phi_k\right\} \\ S(u|o) &\propto \exp\left\{w_S \int P_S(\phi_k|D_k)\cdot U_S(u;o,\phi_k)\,d\phi_k\right\} \end{aligned}$$

This formulation may be interpreted as each agent choosing an action proportional to its expected utility across different possible values of ϕ_k, weighted by the agent's current posterior beliefs about the lexicon their partner is using.

This formulation contrasts with the one suggested by

Bergen et al. (2016), which assumes the expectation takes places at a single level of recursion, say the L_1, as above, and then derives the other agent's behavior by having them reason directly about this marginalized distribution, e.g.

$$\begin{aligned} U_{alt1}(u;o) &= (1-w_C)\cdot \log L(o|u) - w_C \cdot c(u) \\ S_{alt1}(u|o) &\propto \exp\{w_S \cdot U_{alt1}(u;o)\} \end{aligned}$$

where $L(o|u)$ is defined as above. This formulation may be interpreted as an assumption on the part of the speaker that the listener is already accounting for their own uncertainty, and best responding to such a listener. Isolating lexical uncertainty over ϕ to a single level of recursion is a natural formulation for one-shot pragmatic phenomena, where additional layers of recursion can build on top of this marginal distribution to derive implicatures. However, the interpretation is messier for the multi-agent setting, since it (1) induces an asymmetry where one agent considers the other's uncertainty but not vice versa, and (2) requires the speaker to use their own current posterior beliefs to reason about the listener's marginalization.

A third possible variant is to place the expectation outside the listener distribution but inside the speaker's informativity term, i.e..

$$\begin{aligned} L_{avg} &= \int P(\phi_k|D_k)\cdot L_0(o|u,\phi_k)d\phi_k \\ U_{alt2}(u;o) &= (1-w_C)\cdot \log L_{avg}(o|u) - w_C \cdot c(u) \\ S_{alt2}(u|o) &\propto \exp\{w_S \cdot U_{alt2}(u;o)\} \end{aligned}$$

The interpretation here is that the speaker first derives a distribution representing how a listener would respond *on expectation* and then computes their surprisal relative to this composite listener. While this variant is in principle able to derive the desired phenomena, it can be shown that it induces an unintuitive initial bias under a uniform lexical prior, since the logarithm cannot distribute over the integral in the normalization constant. This bias is most apparent in the case of context-sensitivity (Simulation 3).

Mathematically, the difference between these alternatives is whether the speaker's uncertainty about ϕ_k goes inside the renormalization of $L(o|u)$ (as in S_{alt1}), outside the renormalization but inside the logarithm (as in S_{alt2}), or over the entire utility (as in our chosen formulation). While other formulations are conceivable, we argue that marginalizing over the entire utility is not only the most natural but also normatively correct under Bayesian decision theory. When an agent is uncertain about some aspect of the decision problem, rational choice requires the agent to optimize expected utility marginalizing over subjective uncertainty, as in our formulation.

Appendix B: Alternative lexical representations

In this section, we re-consider two specific choices we made about how to represent lexical meanings.

First, for simplicity and consistency with earlier models of Bayesian word learning, we adopted a traditional truth-conditional representation of lexical meaning throughout the paper. Each word in the lexicon is mapped to a single 'concept' to , e.g. w_1 = '*bluesquare*', where this utterance is true of objects the fall in the given concept, and false otherwise. The inference problem over lexicons therefore requires searching over this discrete space of word-concept mappings. However, it is important to emphasize that our model is entirely consistent with alternative lexical representations.

For example, for some settings, a *continuous, real-valued* representation may be preferred, or a higher-dimensional vector representation. Rather than assigning each word a discrete concept in the lexicon, we may simply assign each word-object pair (w_i, o_j) a scalar meaning representing the extent to which word w_i applies to object o_j, such that ϕ is a real-valued matrix:

$$\phi = \begin{bmatrix} \phi^{(11)} & \phi^{(12)} & \dots & \phi^{(1i)} \\ \phi^{(21)} & \phi^{(22)} & \dots & \phi^{(2j)} \\ \vdots & \vdots & \ddots & \vdots \\ \phi^{(j1)} & \phi^{(j2)} & \dots & \phi^{(ij)} \end{bmatrix}$$

and $\mathcal{L}_\phi(w_i, o_j) = \phi^{(ij)}$. In this case, rather than discrete categorical priors over meanings, we may place Gaussian priors over the entries of this matrix:

$$\begin{aligned} \Theta^{(ij)} &\sim \mathcal{N}(0,1) \\ \phi^{(ij)} &\sim \mathcal{N}(\Theta^{(ij)}|1) \end{aligned}$$

We have previously achieved similar results using this alternative lexical representations in earlier iteration of this manuscript (Hawkins et al., 2017; Hawkins, Goodman, et al., 2020), although deriving predictions required variational inference techniques rather than

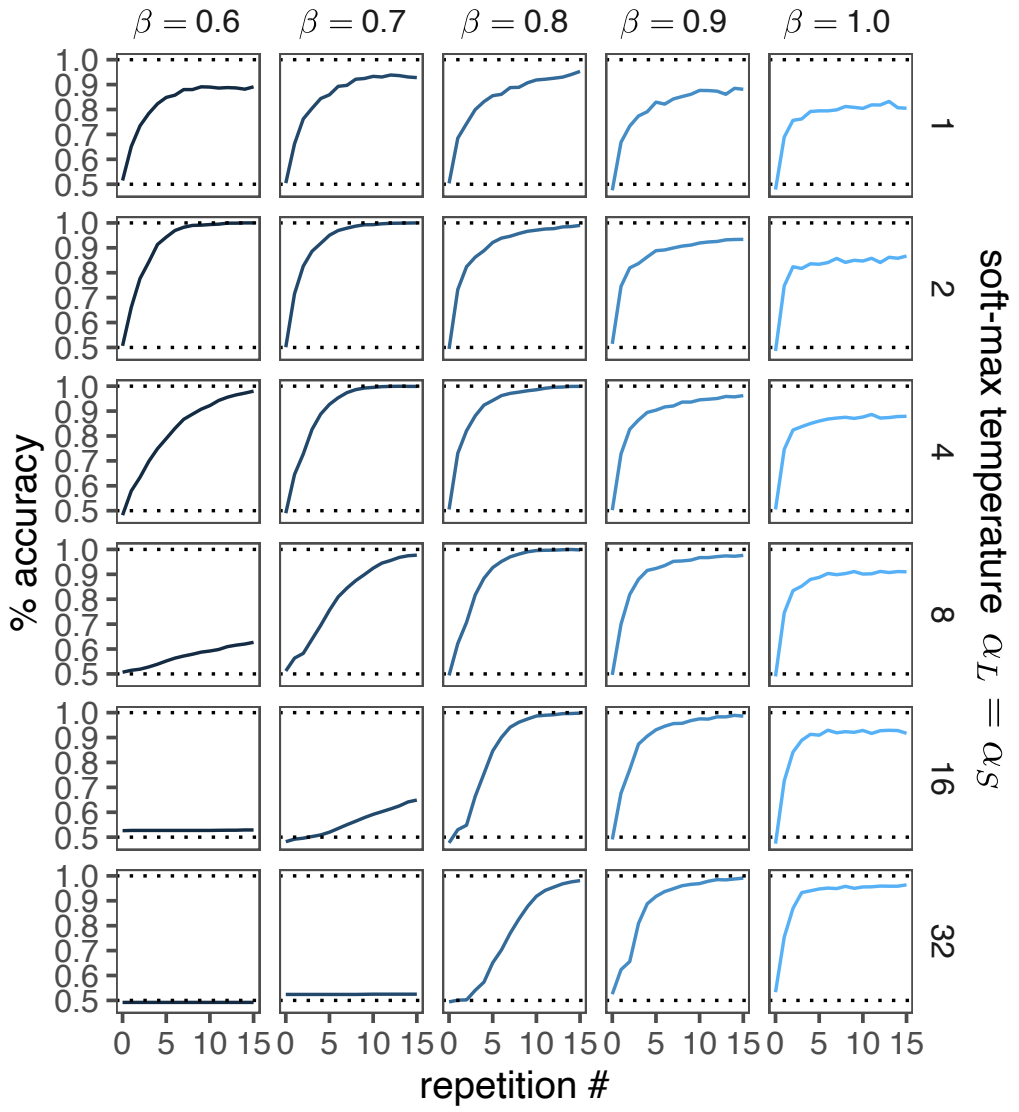

Figure A1. Coordination success (simulation 1.1) across a range of parameter values. Columns represent memory discount parameter β, and rows represent agent soft-max temperatures, where we set $\alpha_S = \alpha_L$. Communicative success is achieved under a wide range of settings, but convergence is limited in some regimes. For example, at high values of β, with no ability to discount prior evidence, accuracy rises quickly but asymptotes below perfect coordination; at low α, inferences are slightly weaker and agent actions are noisier, slowing convergence; finally, at low values of β, when prior evidence is forgotten too quickly, convergence interacts with α: the latest evidence may overwhelm all prior evidence, preventing the accumulation of shared history. The agent noise model is set to $\epsilon = 0.01$ in all simulations.

Markov Chain Monte Carlo. Such optimization-based inference techniques may also provide the most promising path for extending our adaptive model to larger language models, including neural networks that operate over continuous spaces of image pixels and natural language embeddings (Hawkins, Kwon, et al., 2020).

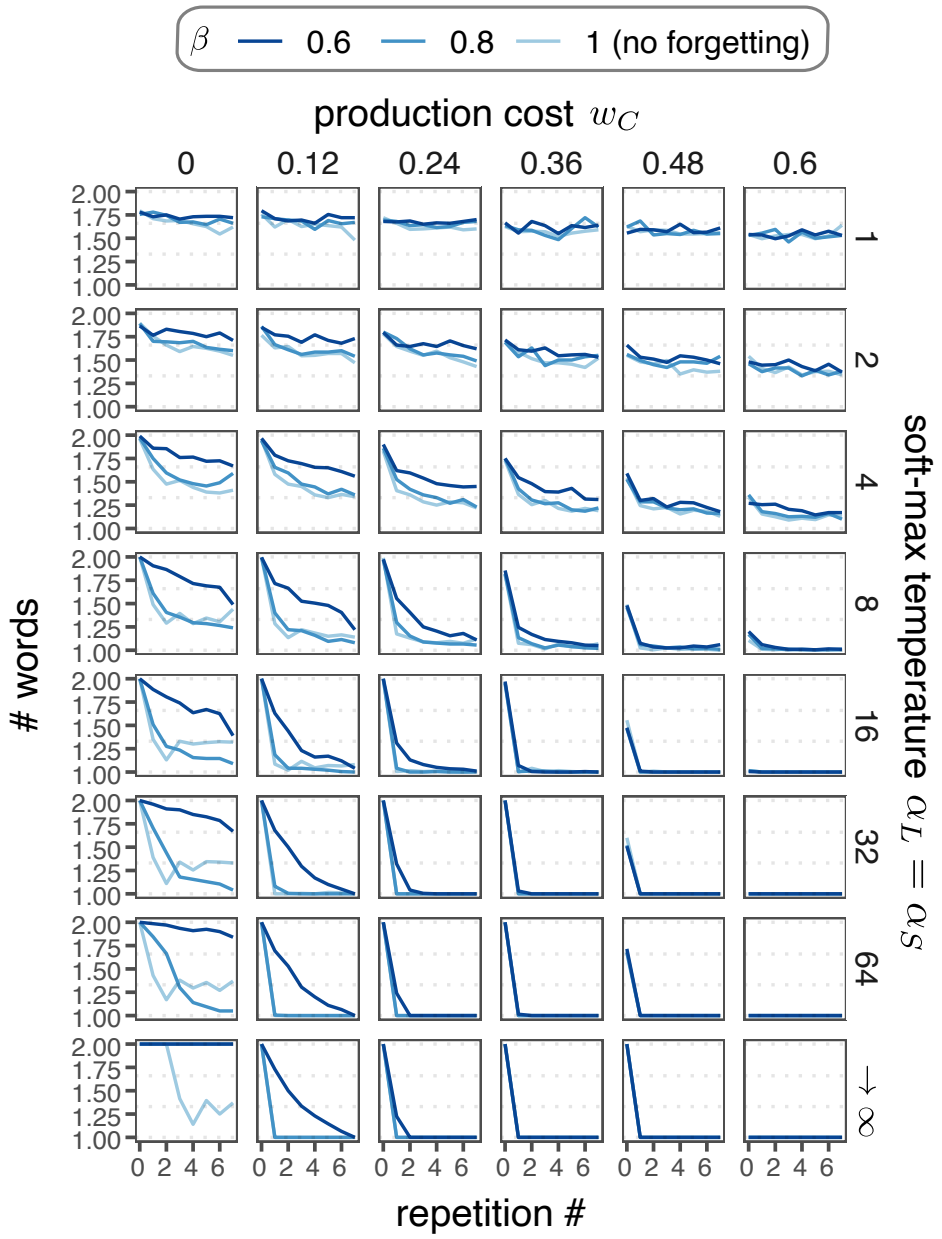

Figure A2. Speaker efficiency (simulation 1.2) across a range of parameter values representing different weights on informativity and cost. Rows represent agent optimality $w_S = w_L$, columns represent costs w_C, and different memory discount factors β shown in different colors. Agents converge on more efficient ad hoc conventions for a wide regime of parameters. When utterance production cost w_C is more heavily weighted relative to informativity, the speaker is less likely to produce longer utterances, even at the beginning of the interaction; when optimality w_S, w_L is higher, and the speaker maximizes utility, we observe faster reduction and more categorical behavior. Note that as $\alpha \to \infty$, utterances only become shorter at $w_C = 0$ in the absence of forgetting. In this case, the shorter utterances approach the exact same utility as the longer utterance, and the speaker reaches equilibrium simply sampling among them at random (i.e. choosing the longer utterance with 1/3 probability and each of the shorter utterances with 1/3 probability).

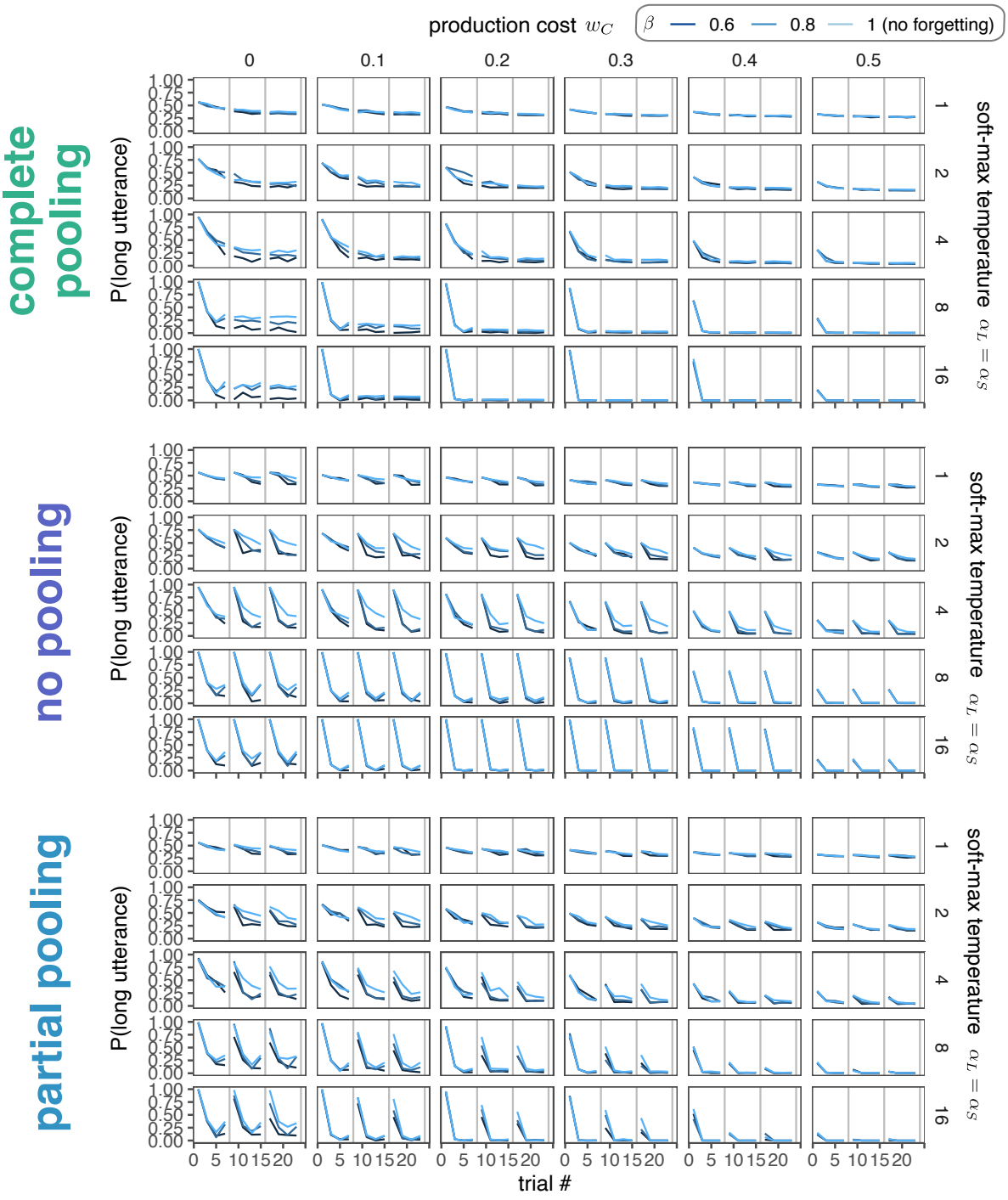

Figure A3. Speaker efficiency simulations for **P2** across a larger parameter regime. We examine the behavior of complete-pooling, no-pooling, and partial-pooling models, where rows represent agent optimality $\alpha_S = \alpha_L$, columns represent cost weight w_c, and colors represent memory discount parameter β.

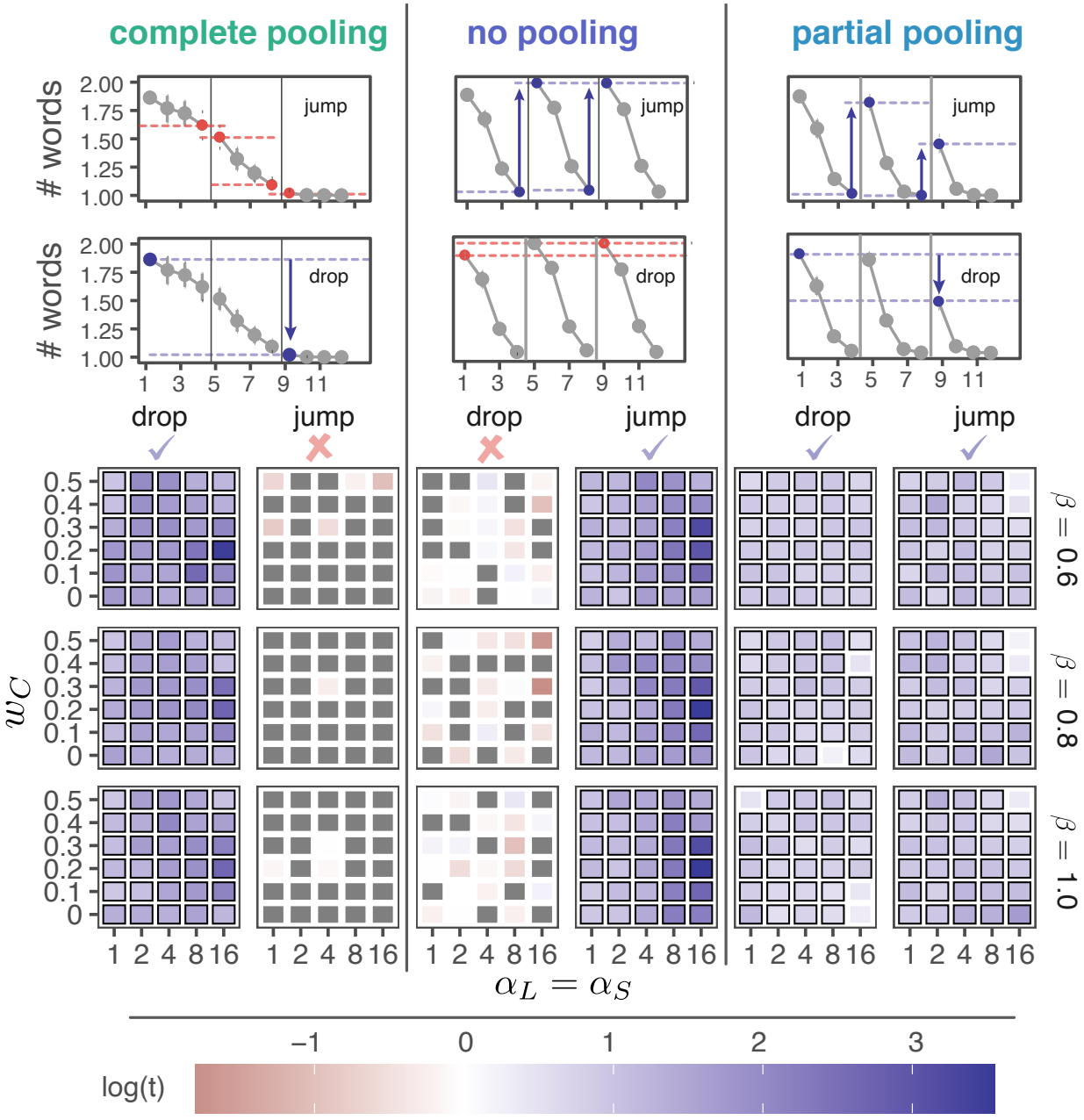

Figure A4. Qualitative predictions of our three models for **P2**. Across a wide range of parameter values, only the hierarchical model consistently produces both qualitative phenomena of interest: reversion to the prior at partner boundaries (i.e. a "jump") and gradual generalization across partners (i.e. a "drop"). Approximately $N = 10$ simulations used to compute t-statistic in each cell. Cells marked with black boxes are significantly different from a null effect of 0 change, $p < 0.005$.

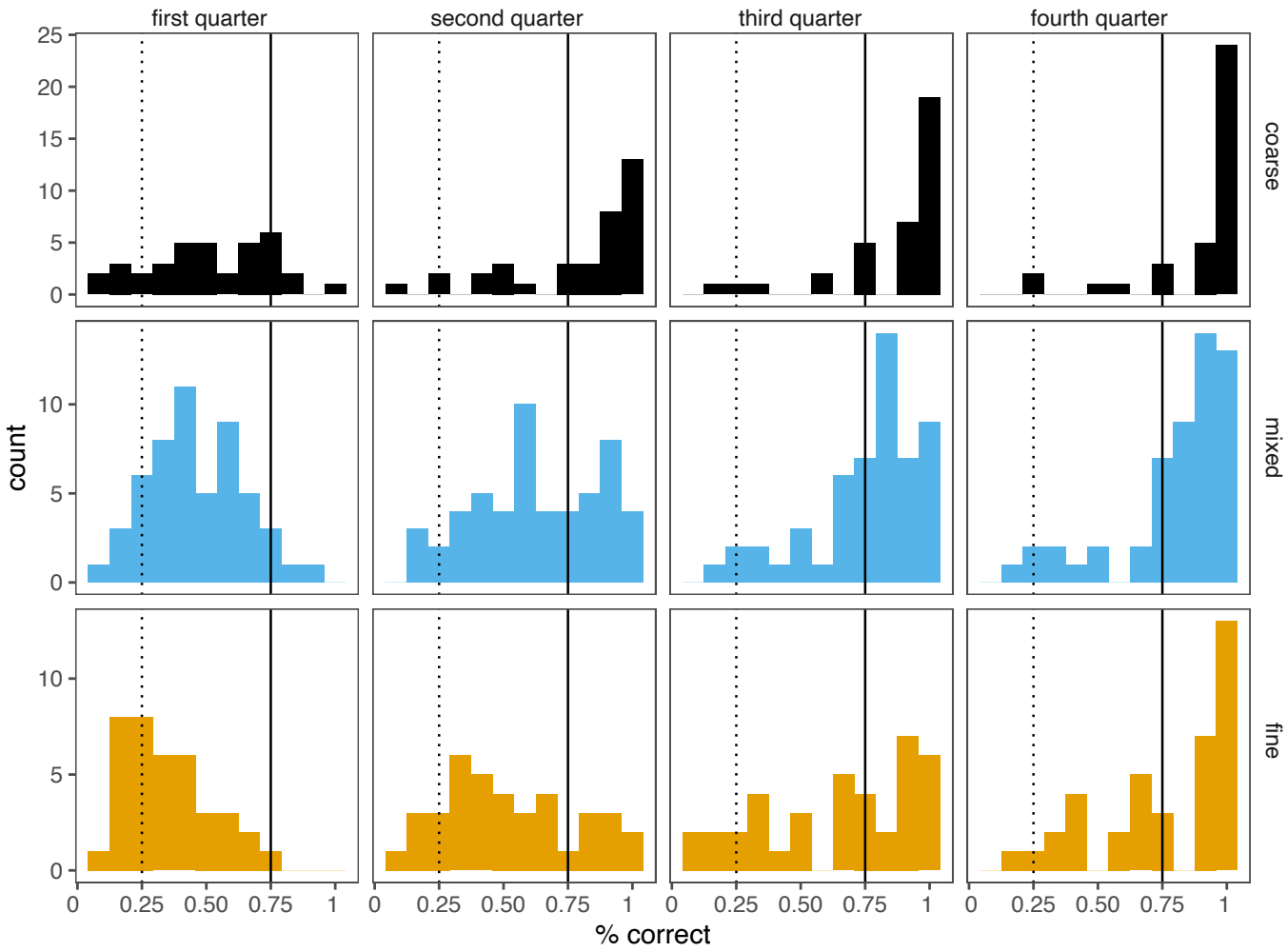

Figure A5. Raw empirical accuracy distributions for context-sensitivity experiment. Each row represents how the accuracies of different games shift across the four quarters of the task for a given condition. Dotted vertical line represents chance accuracy (0.25), solid vertical line represents pre-registered convergence threshold (0.75).

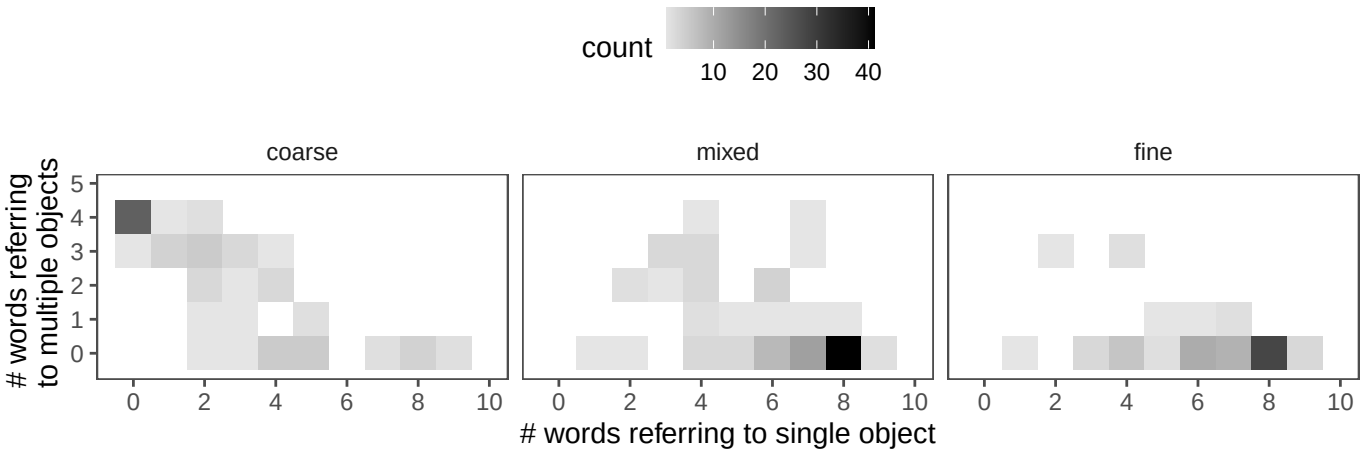

Figure A6. Empirical mixtures of terms reported by participants in **P3**. While the modal lexicon in the coarse condition contained 0 specific terms and 4 more general terms (32% of participants) and the modal lexicon in the mixture and fine conditions contained 8 specific terms and 0 more general terms (42% and 38% of participants, respectively), many participants reported a mixture of abstract and specific terms.