Edinburgh Research Explorer

From partners to populations

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], and 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 CHAI (Continual Hierarchical Adaptation through Inference), a hierarchical Bayesian theory of coordination and convention formation that aims to reconcile the long-standing tension between these two basic observations. We argue that the central computational problem of communication is not simply transmission, as in classical formulations, but continual *learning* and *adaptation* over multiple timescales. Partner-specific common ground quickly emerges from social inferences within dyadic interactions, while community-wide 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 computational foundation for 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 strangers, and (3) the influence of communicative context on which conventions eventually form.

Keywords: communication; learning; convention; inference; generalization; coordination; language; meta-learning

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 expec-

tations extend to complete strangers. An English speaker may order an "espresso" 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, or take on new *ad hoc* senses over the course of a conversation (Clark, 1996). 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.

The tension between these two basic observations, global stability and local flexibility, has posed a challenging 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; Shoham & Tennenholtz, 1997; Barr, 2004; Skyrms, 2010; Steels, 2011; Young, 2015). These accounts typically examine groups of inter-

---

*Correspondence should be addressed to Robert Hawkins, e-mail: rdhawkins@princeton.edu

acting agents who update their representations 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 the development of partner-specific common ground (Clark & Marshall, 1981; Clark & Wilkes-Gibbs, 1986) across extended interactions with the same partner typically do not specify mechanisms by which community-wide conventions 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 background expectations ("prior theories") into interactions, 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 conclude that "there is no such thing as a language" (p. 265), and to abandon appeals to convention altogether (see Heck, 2006; Lepore & Ludwig, 2007; Hacking, 1986; Dummett, 1994 for further discussion of Davidson's view; Armstrong, 2016a, 2016b, provides a philosophical foundation for our synthesis).

In this paper, we propose an account 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:

C1: the ability to represent **variability** about what words will mean to different partners,

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

C3: the ability to gradually *generalize* stable expectations about meaning from individual interactions.

These properties are naturally formalized in a hierarchical Bayesian framework, which we call CHAI (Continual Hierarchical Adaptation through Inference). Indeed, one of our central theoretical aims is to ground the problem of convention formation — 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 coordination and convention:

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

P2: the transition from **partner-specific pacts** to communal conventions 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 CHAI in detail and highlight several important properties of our formulation. The remainder of the paper proceeds through each of the three phenomena (**P1-P3**) in turn. For each phenomenon, we present computational simulations to evaluate how CHAI explains 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 and convention formation in adulthood and domain-generality of discourse processes, as well as several limitations, addressing questions of scalability and incrementality.

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 or concept. 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, 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, unlike typical studies of referring expression generation (van Deemter, 2016; Degen, Hawkins, Graf, Kreiss, & Goodman, 2020; Dale & Reiter, 1995), *repeated reference games* ask speakers to refer to the same targets multiple times as they build up a shared history of interaction with their partners (see Table A1 in Appendix for a review of different axes along which the design has varied). 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 know their partner's identity and maintain the same partner throughout extended interactions. 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 now highlight three findings of particular theoretical significance that emerge from the repeated reference paradigm.

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")[1]. 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). This observation sets up a first puzzle of *ad hoc* convention formation in dyads: How does a word or short description that would be largely ineffective at the outset of a conversation 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)[2]. For example, Wilkes-Gibbs and Clark (1992) paired participants for a stan-
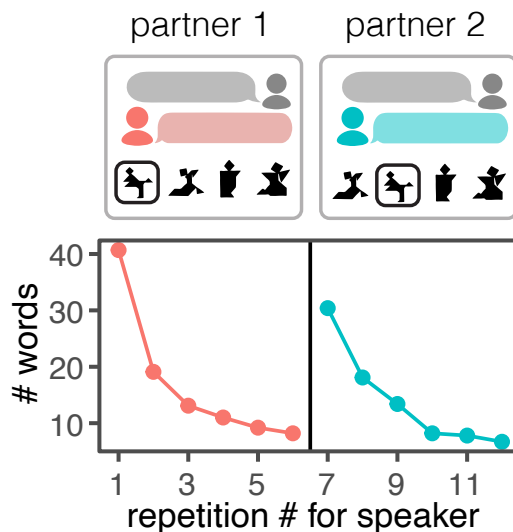

Figure 1

Classic phenomena in repeated reference games. Over multiple iterations with the same partner, speakers speakers converge on increasingly efficient referring expressions (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). Comprehension failures tend to be rare ($\sim$ 2.3%) throughout the experiment, indicating that speakers modulate their utterances effectively. Data from Table 3 in Wilkes-Gibbs and Clark (1992).

dard repeated reference game, but after six rounds, the listener was 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 then coordinated on new

---

[1]Of course, referring expressions are also lengthened for many reasons other than pure reference, such as politeness (*Professor Davidson vs. Don*), affection (*Donny vs Don*), emphasis (*the one and only*), or any number of manner implicatures (see Horn, 1984; Levinson, Stephen, & Levinson, 2000). However, the marked meanings of these longer forms are only obtained against the backdrop of an unmarked or "default" form; repeated reference games set these other functions aside to examine where unmarked expectations come from and how they depend on discourse context (Grosz, 1974; Grosz & Sidner, 1986). This distinction may be seen by considering the non-referential implicatures that may be triggered if a speaker suddenly switched from "martini" back to their original longer description at the end of a game.

[2]We use the term "*ad hoc* convention" (inspired by Barsalou, 1983) interchangeably with the more common "conceptual pact" (Brennan & Clark, 1996; Ibarra & Tanenhaus, 2016) to emphasize the theoretical relationship between this construct and the usual sense of convention referring to longer-term communal knowledge.

ad hoc conventions with their new partner (see Fig. 1). These effects raise our second puzzle: how do community-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, 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) during the first few partner swaps, consistent with partner-specificity, but with subsequent partners, their initial drawings showed a gradual convergence with the conventionalized drawings they had settled upon 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 also 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 collapses across repeated dyadic interactions to exclusively analyze network-level metrics, making it difficult to assess partner-specificity. Given these limitations of existing data, we evaluate our model's predictions using new data from 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 given meaning are equally likely in practice, or even any convention will form at all (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 in individuals may be necessary or sufficient to give rise to such conventions (see Brochhagen, 2021, which raises a similar linking problem).

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

Convention formation as Hierarchical Bayesian inference

In this section, we propose a unified computational account of *ad hoc* coordination and convention formation that aims to address 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 transmission problem has since been enriched to account for *pragmatics* – the ability of speakers and listeners to use context and social knowledge to go beyond the literal meaning of messages (Rosenberg & Cohen, 1966; 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 (see Appendix A for technical details.) In the next section, we review this basic framework and raise two fundamental computational problems facing it. These problems motivate the introduction of continual learning in the CHAI model.

Models of communication with static meaning

For concreteness, we restrict our scope to reference in a context C containing a discrete set of objects $o \in O$, but

the same formulation aims to apply more generally. In this referential setting, 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 $o^* \in C$. They attempt to satisfy Gricean Maxims (Grice, 1975) by selecting utterances according to a utility function $U(u; o)$ that balances informativity to an imagined listener against the cost of producing an utterance. Specifically, S_1 chooses from a "softmax distribution" concentrating mass on the utterance that maximizes $U(u; o)$ to an extent modulated by a free parameter $\alpha_S \in [0, \infty]$:

$$S_1(u|o) \propto \exp\{\alpha_S \cdot U(u; o)\} \tag{1}$$

For $\alpha_S = 1$, this decision rule corresponds to Luce's choice axiom (Luce, 1959). Larger settings of α_S concentrate more probability on the single utterance maximizing utility.

The basic speaker utility function in the RSA framework is defined as follows:

$$U(u; o) = (1 - w_C) \cdot \underbrace{\log L_0(o|u)}_{\text{informativity}} - w_C \cdot \underbrace{c(u)}_{\text{cost}} \tag{2}$$

where $c(u)$ is a function giving the cost of producing u, assuming longer utterances are more costly, and $w_C \in [0, 1]$ is a second free parameter controlling the relative weight of informativity and parsimony in the speaker's production. Critically, the informativity term in Eq. 2 is defined by how well u transmits the intended target o^* to an imagined listener. The simplest imagined listener, L_0, is typically called the "literal listener" because they are assumed to identify the target relying only on the literal meaning of the received utterance, without appealing to further social reasoning about the speaker. That is, the probability of the imagined listener choosing object o is simply assumed to be proportional to the meaning of u under some (static) lexical function $\mathcal{L}$:

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

Throughout this paper, we will take $\mathcal{L}$ to be a traditional Boolean function evaluating whether or not the expression u applies to the entity in question[3]:

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

Two fundamental problems for static meaning

The RSA framework and its extensions provide an account 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 section, we highlight two under-appreciated challenges of communication that complicate this assumption.

The first problem arises from the existence of *variability* throughout a language community (Kidd, Donnelly, & Christiansen, 2018; Wang & Bi, 2021). Different listeners may recover systematically different meanings from the same message, and different speakers may express 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 patients. The words may not be in the patient's lexicon, or common words may be used in non-standard ways. That is, being fluent speakers of the same language does not ensure agreement for the relevant meanings expressed in every context. Different partners may be using different functions $\mathcal{L}$.

The second problem arises from 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 (Gerrig & Gibbs Jr, 1988). 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. *tweeting*, *the cloud*). 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 fixed $\mathcal{L}$ shared by a group of speakers at one moment in time can quickly become outdated (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.

CHAI: A model of dynamic meaning

Rather than assuming a monolithic, universally shared language, we argue that agents solve the fundamental problems posed by variability and non-stationarity by attempt-

_____

[3] Due to the current limitations of representing lexical meaning in formal semantics, it has not been straightforward to specify a truth-conditional function explaining listener behavior for natural-language utterances (e.g. what makes one drawing belong in the literal extension of "upside-down martini glass" but not another, when neither of them are literally martini glasses?) This representation is convenient for our simulations, where we consider all possible discrete mappings between utterances and objects in the context, but better representations of lexical meaning may be substituted (see Potts, 2019). For example, Appendix B works out an example using a real-valued, continuous function (Degen et al., 2020) such as those learned by multi-modal neural networks (Monroe, Hawkins, Goodman, & Potts, 2017; Achlioptas, Fan, Hawkins, Goodman, & Guibas, 2019; Hawkins, Kwon, Sadigh, & Goodman, 2020).

ing to continually, adaptively *infer* the system of meaning used by their current partner. When all agents are continually learning in this way, and changing their own behavior to best respond, we will show that they are not only able to coordinate on local, *ad hoc* meanings or pacts with specific partners but also abstract away *conventions* that are expected to be shared across an entire community. We introduce the CHAI (Continual Hierarchical Adaptation through Inference) model in three steps, corresponding to how it formalizes the three core capacities **C1-C3**: hierarchical uncertainty about meaning, online partner-specific learning, and inductive generalization.

C1: Representing variability in meaning via structured uncertainty. 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 monolithic, 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. 2 can be re-written to give behavior under a fixed value of ϕ:

$$L_0(o|u,\phi) \propto \mathcal{L}_\phi(u,o) \qquad (3)$$
$$U(u;o,\phi) = (1 - w_C) \cdot \log L_0(o|u,\phi) - w_C \cdot c(u)$$
$$S_1(u|o,\phi) \propto \exp\{\alpha_S \cdot U(u;o,\phi)\}$$

While we will remain agnostic for now to the exact functional form of $\mathcal{L}_\phi$ and the exact parameter space of ϕ, there are two 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*. That is, an agent should know that more people will share the meaning of some words than others, and should conversely expect more consensus about how to refer to some concepts than others. Second, expectations about which meanings will be evoked for a given utterance and which utterances are expected to be used to express a meaning should be *sensitive to the social identity of one's partner*.

The first desideratum – the ability to represent variability in the population – 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 each word as used by different partners. In a Bayesian framework, this uncertainty is specified by a prior probability distribution $P(\phi)$ over possible function parameters. For example, imagine a doctor giving a diagnosis to a new patient. Under some possible values of ϕ, a piece of medical jargon like "sclerotic aorta" refers unambiguously to the patient's heart condition. Under other values of ϕ, it has a less clear meaning. A doctor with good bedside manner should assign some

probability to each possibility rather than assuming everyone will share the same precise meaning they learned in medical school. Importantly, this variability will be different for different words: likely more people share the meaning of "dog" than "sclerotic aorta". This core idea of introducing uncertainty over a partner's lexical semantics has previously been explored in the context of one-shot pragmatic reasoning, where it was termed *lexical uncertainty* (Bergen, Goodman, & Levy, 2012; Potts & Levy, 2015; Bergen, Levy, & Goodman, 2016; Potts, Lassiter, Levy, & Frank, 2016), as well as in the context of iterated dyadic interactions (Smith, Goodman, & Frank, 2013).

Second, this representation should also, in principle, be sensitive to the social identity of the partner: a doctor should be able to form different expectations about a new colleague than a new patient (Clark, 1998). This 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 "overhypothesis" about the overall distribution of all possible partners. This level can be viewed as a representation of long-term "communal lexicons" about common knowledge based on community membership (Clark & Marshall, 1981). 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). ϕ_k can be viewed as the "idiolect" that has been fine-tuned to account for partner-specific common ground and conceptual pacts from previous interactions[4].

──────────

[4]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

shared representation   partner 1 Θ partner k

lexical prior for
individual partner $P(\phi_1|\Theta) \quad \cdots \quad P(\phi_k|\Theta)$

partner-specific
predictions $P(D_1|\phi_1) \ldots P(D_k|\phi_k)$
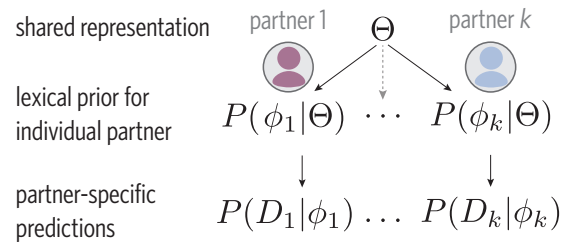
Figure 2

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

To integrate 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.

$$L(o|u) \propto \exp\left\{\alpha_L \cdot \int P_L(\phi_k) \log S_1(u|o, \phi_k) \, d\phi_k\right\}$$
$$S(u|o) \propto \exp\left\{\alpha_S \cdot \int P_S(\phi_k) U(u; o, \phi_k) \, d\phi_k\right\} \quad (4)$$

where $\alpha_S, \alpha_L \in [0, \infty]$ control the speaker's and listener's soft-max optimality, respectively[5].

C2: Online learning via partner-specific inference. The formulation in Eq. 4 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:

$$\begin{aligned} P(\phi_k, \Theta|D_k) &\propto& P(D_k|\phi_k, \Theta)P(\phi_k, \Theta) \\ &=& P(D_k|\phi_k)P(\phi_k|\Theta)P(\Theta) \end{aligned} \quad (5)$$

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 of meaning.

Importantly, the posterior obtained in Eq. 5 allows agents to explicitly maintain *partner-specific expectations*, as used in Eq. 4, by marginalizing over community-level uncertainty:

$$P(\phi_k|D_k) = \int_\Theta P(\phi_k, \Theta|D_k) d\Theta \quad (6)$$

We will show that when agents learn about their partner in this way, and adjust their own production or comprehension accordingly (i.e. Eq. 4), they are able to coordinate on stable *ad hoc* conventions.

C3: Generalization to new partners via hierarchical induction. The posterior in Eq. 5 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:

$$P(\Theta|D) = \int_\phi P(\phi, \Theta|D) d\phi \quad (7)$$

where $D = \bigcup_{k=1}^N D_k$, $\phi = \phi_1 \times \cdots \times \phi_N$, and N is the number of partners previously 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 novel partners 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). One consequence 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).

Further challenges for convention formation

The formulation in the previous section presents the core of CHAI. Here, we highlight several additional features addressing more specific challenges raised by prior work on

----

partners belong to distinct sub-communities (e.g. represented by discrete latent variables Gershman, Pouncy, & Gweon, 2017; Gershman & Cikara, 2020), which may explain code-switching (Auer, 2013; Hawkins, Liu, Goldberg, & Griffiths, 2021) and other social inferences based on language use (Kinzler, 2021; Isaacs & Clark, 1987; Roberts, 2010).

[5]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 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 (see also Zaslavsky, Hu, & Levy, 2020).

communication and which we will encounter in the simulations reported in the remainder of the paper. Our organization of these details is motivated by the analysis of Spike, Stadler, Kirby, and Smith (2017), who highlighted three common issues that all accounts of convention formation must address: (1) the form of feedback available, (2) the influence of memory and temporal discounting, and (3) the form of pragmatic reasoning being used. Finally, we set up the basic simulation framework that will be used throughout the rest of the paper.

The role of social observation. Learning and adaptation depend critically on the availability and quality of social observations D_k (Eq. 5). If the speaker has no way of probing the listener's understanding, or if the listener has no way of comparing their interpretation against the speaker's intentions, however indirectly, they can only continue to rely on their prior expectations, with no ground for conventions to form (Hupet & Chantraine, 1992; Garrod, Fay, Lee, Oberlander, & MacLeod, 2007). Communication is empirically hindered under degraded observation conditions (Krauss & Weinheimer, 1966; Krauss & Bricker, 1967; Krauss, Garlock, Bricker, & McMahon, 1977; Schober & Clark, 1989), and we have all been in situations where we thought we were on the same page with a partner and only realized that we misunderstood much later, when the consequences because clear. In principle, we expect that D_k should reflect all relevant sources of information that may expose an agent's state of understanding or misunderstanding. Not just ostensive signals like pointing (van de Braak, Dingemanse, Toni, van Rooij, & Blokpoel, 2021), but verbal and non-verbal backchannels (e.g. *mmhmm*, nods or quizzical looks), forms of self-initiated and other-initiated repair (e.g. clarification questions or requests for confirmation Schegloff, Jefferson, & Sacks, 1977; Dingemanse et al., 2015; Arkel, Woensdregt, Dingemanse, & Blokpoel, 2020), and downstream actions taken in the world (e.g. attempts to follow instructions).

While incorporating these richer sources of information presents an exciting line of future work, we restrict our scope to the feedback traditionally provided by the empirical repeated reference 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. To specify the likelihoods in Eq. 5 for this referential setting, we assume each agent should infer their partner's lexicon ϕ_k by conditioning on their *partner's* previous behavior. 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[6]. 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.

The role of memory and recency. One important constraint is imposed by the basic cognitive mechanisms of memory. It is unrealistic to expect that memory traces 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 a mechanism by which errors become less accessible, early misunderstandings may interfere with coordination much later in an interaction. One possible solution is to privilege more recent outcomes. Especially if a partner is assumed to change over time, then older data may provide less reliable cues to their current behavior. Recency is typically incorporated into Bayesian models with a simple 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 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 auto-correlation between neighboring time points yields this form of discounting[7].

---

[6] 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 (Misyak, Melkonyan, Zeitoun, & Chater, 2014), but we leave consideration of such asymmetries for future work.

[7] While this simple decay model is sufficient for our reference games, it is clearly 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 or source memory. It is also con-

The role of pragmatics. While natural languages are rife with ambiguous and polysemous terms, speaker and listeners must somehow resolve these ambiguities to be understood in context (Piantadosi, Tily, & Gibson, 2012)[8]. For example, Brennan and Clark (1996) placed participants in a context where the target object o^* was easily distinguished from other objects in the context C by a referring expression like u ="the shoe." In a second phase of the study, the context C' was switched to be a set of other shoes. Even though there was strong precedent for referring to o^* as "the shoe," this description was no longer *informative*: the speaker recognized that u could apply equally well to all $o \in C$ leading to potential ambiguity about which shoe they were referring to. As a result, the speaker switched to a more specific utterance like u' ="the pennyloafer" which unambiguously applied to o^* in the new context. In a third and final phase, the context reverted back to the original one, C, but many speakers continued to use the more specific utterance u' even though u would have sufficed. This example emphasizes how *ad hoc* conventions or pacts may be sensitive to the context in which they form.

CHAI solves this problem by the principles of *pragmatic reasoning* naturally instantiated in the RSA framework (Frank & Goodman, 2012), which plays two distinct roles. First, our Gricean agents' production and comprehension is guided by cooperative principles (Eq. 4). They do not only make passive inferences from observation, they participate in the interaction by *using language* themselves. Second, our agents assume that their *partner* is also using language in a cooperative manner, which strengthens the inferences they may make about the underlying system of meanings their partner is using. That is, we use the RSA equations as the linking function 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. 5). This use of pragmatic reasoning has been explicitly linked to principles like mutual exclusivity in word learning (Bloom, 2002; Frank, Goodman, & Tenenbaum, 2009; Smith et al., 2013; Gulordava, Brochhagen, & Boleda, 2020; Ohmer, König, & Franke, 2020). For example, upon hearing their partner use a particular utterance u to refer to an object o, a pragmatic listener can not only infer that u means o in their partner's lexicon, but also that other utterances u' likely do *not* mean o: if they did, the speaker would have used them instead.

Simulation 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 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, consider a simple reference game with two utterances and two objects ($o_1 = \blacksquare$ and $o_2 = \bullet$). Then there are four possible lexicons, corresponding to the four assignments of objects to utterances:

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

Second, having defined the support of the parameter ϕ, we can then define a lexical prior. We consider a partition-based simplicity prior based on the size of the lexicon (Frank et al., 2009; Carr, Smith, Culbertson, & Kirby, 2020): $P(\phi) \propto \exp\{-|\phi|\}$, where $|\phi|$ is the number of lexical items. 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:

$$P(\phi) = \begin{bmatrix} \text{Unif}\{\blacksquare, \bullet\} \\ \text{Unif}\{\blacksquare, \bullet\} \end{bmatrix} = \begin{matrix} \blacksquare \quad \bullet \\ \begin{bmatrix} .5 & .5 \\ .5 & .5 \end{bmatrix} \end{matrix} \begin{matrix} u_1 \\ u_2 \end{matrix}$$

This simplicity prior becomes more important for **P3**, where we consider spaces of referents with more complex conceptual structure. A single word may apply to multiple conceptually related referents (e.g. all of the squares) or, conversely, may apply to no referents at all, in which case it is effectively removed from the agent's vocabulary. In this case, the simplest lexicon is a single word that refers to everything and the most complex lexicon assigns a unique word for each object (see Appendix C for discussion of alternatives.)

---

sistent with multiple algorithmic-level mechanisms; for example, decay can be viewed as a form of weighted importance sampling, where more recent observations are preferentially sampled (Pearl, Goldwater, & Steyvers, 2010), or a process where observations have some probability of dropping out of memory at each time step.

[8]Indeed, Brochhagen (2020) has suggested that high degrees of lexical ambiguity and polysemy, i.e. high degrees of uncertainty over Θ in CHAI, are useful precisely because they allow much-needed flexibility supporting partner-specific adaptation.

Finally, while the probabilistic model we have formulated in this section is theoretically well-motivated and mathematically well-defined, it is challenging to derive predictions from it. Historically, interactive models like ours are not amenable to 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 (see Appendix A for further details of our implementation.)

Summary

In this section, we formalized the computational problem facing agents who must communicate in a variable, changing world. No static lexicon is appropriate for all partners and situations, requiring them to update on the fly. We proposed CHAI, a cognitive model of how people solve this problem through continual adaptation. CHAI instantiates three core capacities in a hierarchical Bayesian framework: (C1) structured uncertainty over what words mean to different partners, (C2) social inference to back out likely latent systems of meaning from a partner's observable behavior, and (C3) hierarchical induction to generalize to the overall distribution of possible partners. In the remainder of the paper, we argue that CHAI provides a new computational foundation for understanding coordination and convention formation, focusing on three empirical phenomena that have posed a challenge for previous accounts: (P1) the increase in communicative efficiency as a function of shared history, (P2) the transfer of partner-specific expectations to communal expectations, and (P3) the influence of communicative context on which conventions eventually form.

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. In this section, we argue that CHAI provides a rational explanation for increasing efficiency in terms of the inferences made by speakers across repeated 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. 4), 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, $O = \{\bullet, \blacksquare\}$, 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 $\bullet$. 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 $u \sim S(u|\bullet)$. 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, $\bullet$, a correct response. Both agents may use the resulting tuple $D = \{\bullet^*, u_1, \bullet\}$, depicted in the top row in Fig. 3 to update their beliefs about the lexicon their partner is using.

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

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
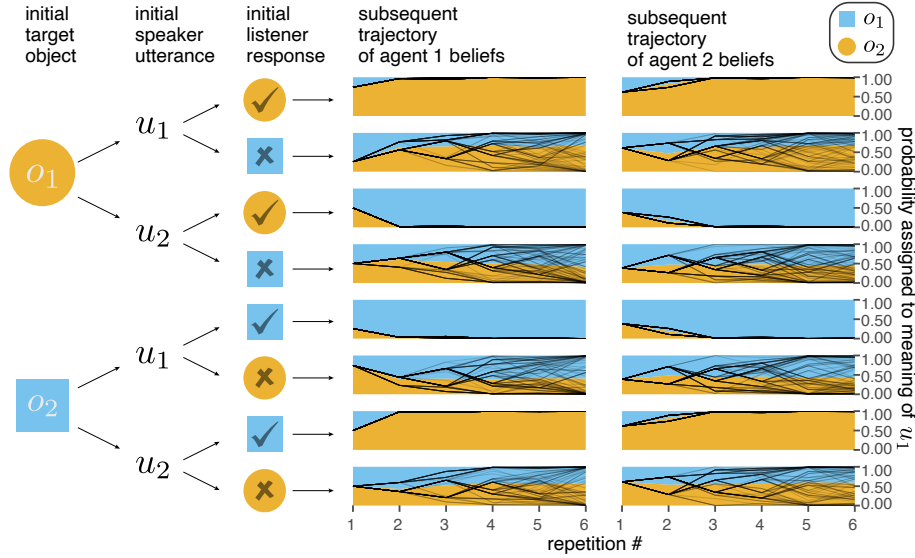
Figure 3

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.)*
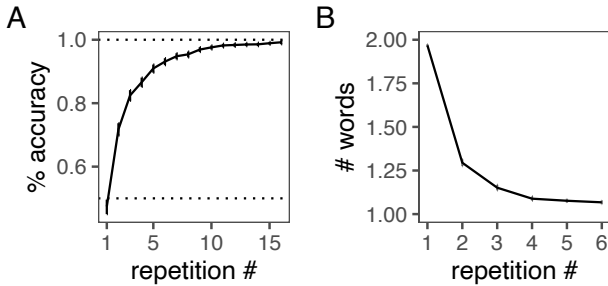

Figure 4

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.*

the beginning of each block. We show representative behavior at soft-max 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. A2).

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 $D = \{(\bullet^*, u_2, \bullet)\}$ 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:

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

One subtle consequence of a conjunctive Boolean semantics is the possibility of contradictions. For example, under a possible lexicon where $\phi(u_1) = \blacksquare$ and $\phi(u_2) = \bullet$, the multi-word utterance u_1u_2 is not only false of the particular referents in the current context, it is false of all *possible* referents, reflecting a so-called truth-gap (Strawson, 1950; Van Fraassen, 1966). We assume such an utterance is uninterpretable and simply disregarded without changing 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 truth-conditional semantics (Degen et al., 2020) and show in Appendix B that switching to a continuous semantics with lexical values in the interval [0, 1] may better capture the notion of redundancy that motivates speakers to initially produce longer utterances.

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):

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

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 ≈ 1.5 on first block) but rapidly converges to shorter utterances after several repetitions (mean length ≈ 1 on final block), qualita-

tively 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. 5). Consider a speaker who wants to refer to object $\blacksquare$. 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 $\bullet$, 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 $\phi(u_1) = \phi(u_2) = \blacksquare$ in the listener's lexicon ($p = 0.55^2 \approx 0.3$), in which case the listener will correctly identify $\blacksquare$ with high probability. The possibility that both $\phi(u_1) = \phi(u_2) = \bullet$ in the listener's lexicon is only $p = 0.45^2 \approx 0.2$, in which case the listener will erroneously select $\bullet$. In the mixed cases, where $\phi(u_1) = \bullet, \phi(u_2) = \blacksquare$ or $\phi(u_1) = \blacksquare, \phi(u_2) = \bullet$ 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, $\log(\epsilon)$ 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, $\blacksquare$). This allows both agents to become more confident that the component utterances u_1 and u_2 mean $\blacksquare$ 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 $\blacksquare$ would have been more likely to say u_1u_2 than a speaker for whom either component meant $\bullet$; and similarly for the speaker reasoning about possible listeners. Consequently, the probability of both mappings increases.

Fig. 5 shows the trajectories of internal components of the speaker utility as the interaction continues. We assume for illustrative purposes in this example that $\blacksquare$ 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 $\blacksquare$ 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
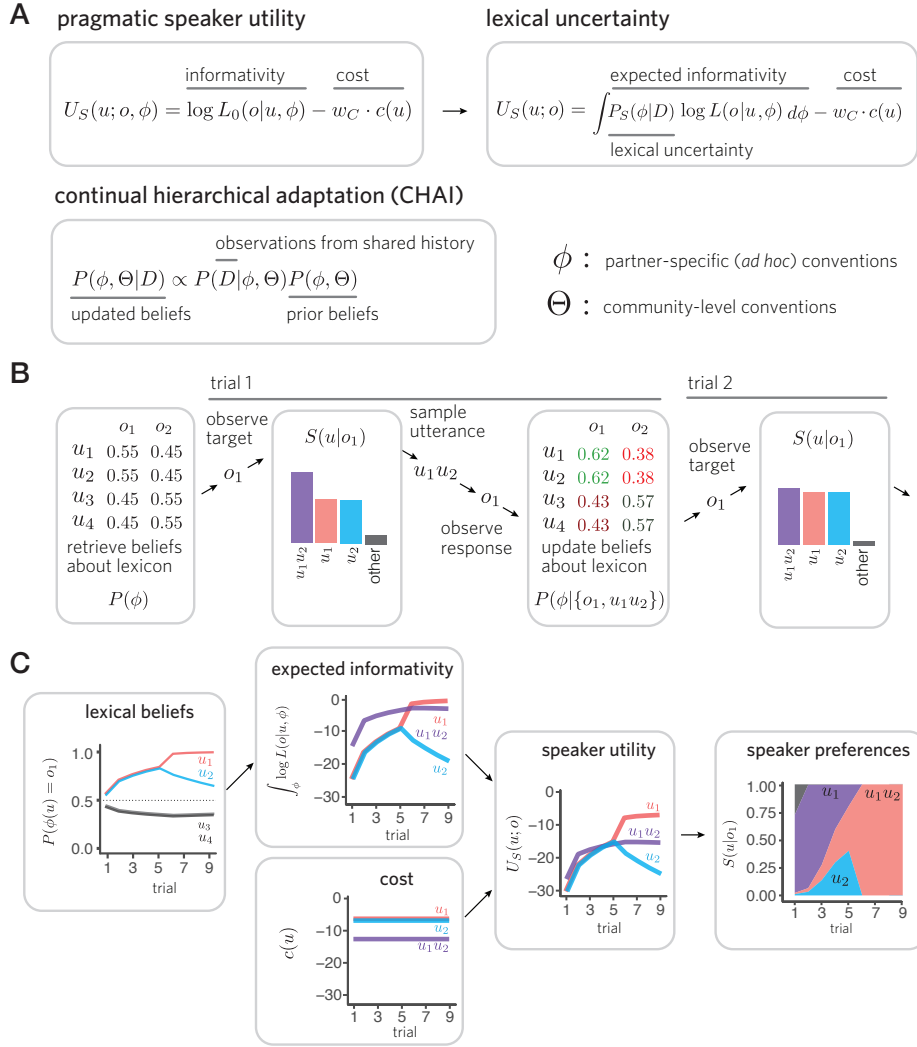
Figure 5

Internal state of speaker. *(A) Schematic showing how lexical uncertainty is added to a simple pragmatic speaker utility; CHAI proposes that lexical expectations are adapted over time based on social observations. (B) A single trial of Simulation 2.1. 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. (C) Each internal term of the speaker's utility (Eq. 4) is shown throughout the 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.*

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 ra-

tional 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. A3).

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. This explanation recalls classic observations about *hedges* (expressions like *sort of* or morphemes like *-ish*) that explicitly mark provisionality, such as *a sort of silvery purple colored car* (Lakoff, 1975; 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 include hedges, it is possible to understand this behavior as an explicit or implicit marker of the lexical uncertainty construct in our account. Our account is also broadly 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 would be expected by random and independent corruption. In other words, speakers apparently begin by combining multiple descriptive modifiers and collapse to retain only one of these 'units' contingent on evidence that their partner understands.

Why has this phenomenon remained outside the explanatory scope of previous models? Our account differs in both level of analysis and model complexity. For example, the influential *interactive alignment* account proposes that that speakers adapt and coordinate on meaning through priming mechanisms that allow phonetic or syntactic features associated with lexical items to percolate up to strengthen higher levels of representation (Pickering & Garrod, 2004, 2006; Roelofs, 1992). While priming mechanisms are certainly at play in repeated reference tasks, especially when listeners engage in extensive dialogue and alternate roles, it is not clear why priming alone would lead to convergence on more efficient descriptions as opposed to aligning on the same longer initial description. Furthermore, priming cannot explain why speakers still converge to shorter descriptions even when the listener is prevented from saying anything at all and only sparse, non-verbal feedback of success is provided, or why speakers continue using longer descriptions when they receive non-verbal feedback that the listener is repeatedly making errors (Krauss & Weinheimer, 1966; see also Hawkins, Frank, & Goodman, 2020). In these cases, there are no linguistic features available for priming or alignment to act upon. To be clear, our computational-level account is not mutually exclusive with these process-level principles and does not in any way falsify or undermine them. Explaining when and why speakers believe that shorter descriptions will

suffice, and how it depends on context, requires additional computational-level principles, which we hope will lead to further enrichment of algorithms at the process level.

Another prominent account proposes that speakers coordinate on meaning using a simpler update rule that simpler makes utterances more likely to be produced after communicative successes and less likely after communicative failures. This account has often been implemented using a simple variant of reinforcement learning (RL) such as Roth-Erev learning (Erev & Roth, 1998; Steels, 1995; Barr, 2004; Young, 2015). While such minimal rules allow groups to reach consensus, it is challenging to explain the full suite of phenomena we have explored in this section. First, it is not clear how simply reinforcing longer descriptions could lead them to get shorter. In the rare cases that have allowed longer utterances to be constructed compositionally from more primitive utterances, reduction has been hard-coded as a kind of ϵ-greedy exploration where the speaker has a fixed probability of dropping a random token at each point in time (Beuls & Steels, 2013; Steels, 2016). Such noisy dropping, however, is inconsistent with studies by Hupet and Chantraine (1992) where 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 would be the same as the interactive task. No evidence of reduction was found in this case, and in some cases utterances actually grew longer (see also Garrod et al., 2007). Even if we fixed this problem by extending the update rule to be contingent on interaction, it is not clear why a speaker would initially prefer to produce longer utterances over shorter utterances.

Importantly, these limitations do not stem from the RL framework itself, but from the simplifying assumption that the probability of taking actions should be directly tied to the previous outcomes of those actions. CHAI preserves a core idea from these accounts — the ability to dynamically adapt one's behavior contingent on one's partner's — but disentangles the inference problem (i.e. estimating a partner's underlying lexicon) from the decision problem (i.e. deciding which action to take with these estimates in hand). Introducing the latent variable of the lexicon increases the model's complexity but is also more explanatory, as we show in the subsequent sections. Importantly, more sophisticated model-based reinforcement learning algorithms make a similar distinction and may consequently be flexible enough to account for this phenomenon (see Gershman & Niv, 2015 for an explicit connection between hierarchical Bayes and an RL algorithm known as TD-learning; but see Vélez & Gweon, 2021 for outstanding problems associated with bridging these perspectives).

Finally, while our simulations captured several core features of the reduction phenomenon, they have only scratched the surface of its empirical complexity. First, our simu-

lations only consider two-word descriptions with homogenous uncertainty over the components, while the semantic components of real initial descriptions have more heterogeneity. It remains an open question as to how best to instantiate more realistic priors in our model that can predict more fine-grained patterns. For example, early hand-tagged analyses by Carroll (1980) 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. Second, gains in efficiency associated with *ad hoc* conventions do not necessarily translate into shorter utterances. Outside of the domain of reference games, speakers often have control over what they want to convey and may use the efficiency afforded by their new conventions to express more information in the same number of words rather than the same amount of information in fewer words (Effenberger, Yan, Singh, Suhr, & Artzi, 2021). Once a convention is formed, it can be used as a new primitive to bootstrap further conventions and convey ever-more-sophisticated meanings (McCarthy, Hawkins, Wang, Holdaway, & Fan, 2021).

Phenomenon #2: Conventions gradually generalize to new partners in community

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 (see Smith et al., 2017, which contrasted no-pooling and complete-pooling models). 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.

In other words, complete-pooling and no-pooling models are *prima facie* unable to explain partner-specificity and network convergence, respectively. CHAI is a hierarchical *partial-pooling* account that offers a solution to this puzzle. We propose that social 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. 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. We then explicitly compare CHAI to complete-pooling and no-pooling sub-models that lesion the hierarchy, using only the top level or bottom level, to evaluate the contribution of each component.

Model predictions: Simulation 2.1

We first examine the generalization behavior produced by each model by simulating the outcomes of interacting with multiple partners on a small network (see Fig. 6A). We used a round-robin scheme (Fig. 6B) 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 within a given interaction. Consequently, all agents at a particular phase have interacted with the same number of previous partners, allowing us to exam-
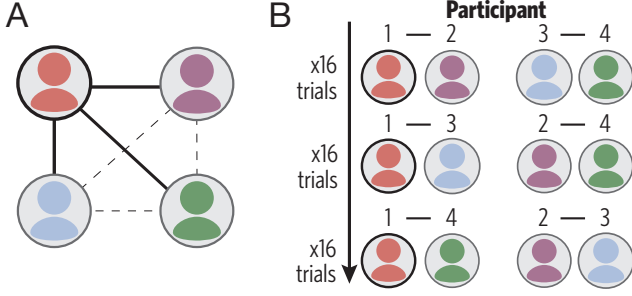
Figure 6

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.

ine network convergence (but see Hawkins, Goodman, et al., 2020, for a "first-person" version where each new partner is entirely fresh to the task, finding similar speaker generalization).

Unlike our previous simulations with a single partner, where hierarchical generalization was irrelevant, we must now specify the hyper-prior P(Θ) governing the overall distribution of *partners* (Eq. 5). Following Kemp et al. (2007), we extend the uniform categorical prior over possible referents to a hierarchical Dirichlet-Multinomial model (Gelman et al., 2014), where the 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 across referents, we assume it is drawn from a Dirichlet prior:

$$\begin{aligned}
\phi_k(u) &\sim \text{Categorical}(\Theta) \\
\Theta &\sim \text{Dirichlet}(\lambda \cdot \boldsymbol{\alpha})
\end{aligned} \quad (8)$$

where $\lambda \cdot \boldsymbol{\alpha}$ gives 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 the scaling factor λ roughly corresponds to prior beliefs about the spread, where larger magnitudes correspond to more concentrated probability distributions across the population. We fix $\lambda = 2$ and assume the agent has uncertainty about the population-level central tendency by placing a hyper-prior on $\boldsymbol{\alpha}$ (see Cowans, 2004) that roughly corresponds to the weak initial preferences we used in our previous simulations:

$$\boldsymbol{\alpha} \sim \begin{cases} \text{Dirichlet}(1.0, 1.5) & \text{if} \quad u \in \{u_1, u_2\} \\ \text{Dirichlet}(1.5, 1.0) & \text{if} \quad u \in \{u_3, u_4\} \end{cases}$$

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 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. A4 in the Appendix for an exploration of other parameters).

Speaker utterance length across partners. We begin by examining our model's predictions about how a speaker's referring expressions 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. 7A, first column). No-pooling accounts predict that roughly the same initial description length will re-occur with every subsequent interlocutor (Fig. 7A, second column).

Here we show that a partial pooling account predicts 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. 7A, third column).

It is possible, however, that these two predictions only distinguish our partial-pooling model 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
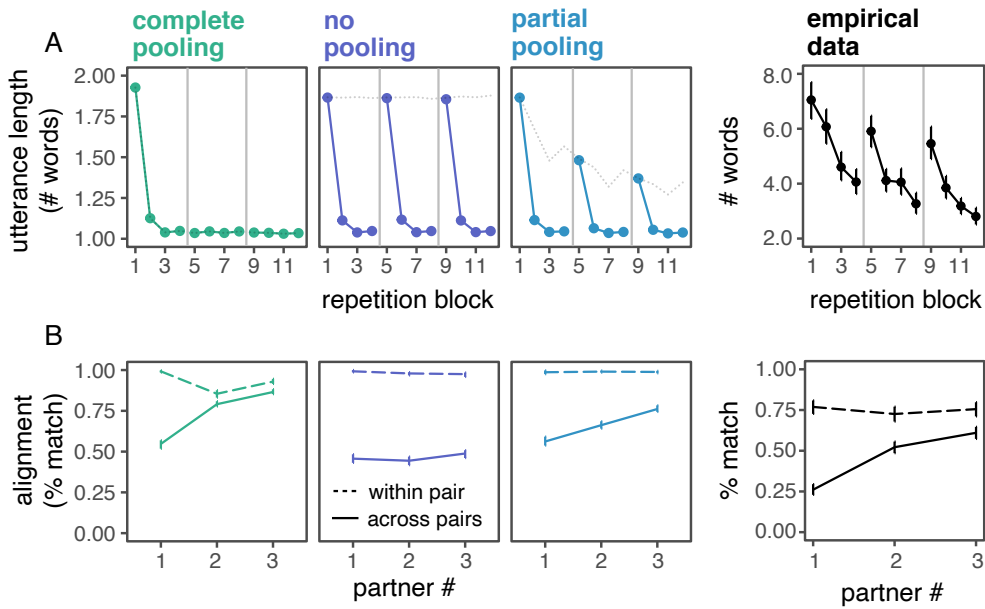
Figure 7

Simulation results and empirical data for (A) speaker reduction, and (B) network convergence across three partners. In (A) vertical boundaries mark time points when new partners were introduced, and the dotted grey line represents what would be produced for a stranger at each point in time. In (B), dashed line represents alignment between partners who are currently paired while solid line represents alignment across partners who are not *paired. Error bars represent bootstrapped 95% confidence intervals.*

model fails to predict generalization nearly everywhere. Detailed results are shown in Fig. A5 in the Appendix.

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. We calculated alignment between currently interacting agents (i.e. *within* a dyad) and those who were not interacting (i.e. *across* dyads), averaging across the target objects. Alignment across dyads was initially near chance, reflecting the arbitrariness of whether speakers reduce to u_1 or u_2. Under a complete-pooling model (Fig. 7B, first column), agents sometimes persist with mis-calibrated expectations learned from previous partners rather than adapting to their new partner, and *within-dyad* alignment deteriorates, reflected by a sharp drop from 99% to 85%. Under a no-pooling model (Fig. 7B, second column), convergence on subsequent blocks remains near chance, as conventions need to be re-negotiated from scratch. 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. 7B, third column).

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 partners (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. 6A). 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. 6B). 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. 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. 6 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 metrics we used in our simulations: utterance length and network convergence.

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 efficiency, 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. 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. 7A, final column), 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.

Network convergence. Now, 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 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: the partial-pooling model predicted that within-pair alignment should stay consistently high while (tacit) alignment between non-interacting pairs will increase. To test this prediction, we 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 signif-

icant interaction ($b = -0.85, z = -5.69, p < 0.001$; see Fig. 7B, final column). Although different pairs in a network may initially use different labels, these labels begin to align over subsequent interactions.

This finding is consistent with the primary prediction of interest for both the complete-pooling and partial-pooling model. These two models only pull apart for a secondary prediction concerning the transition from the first to second partner. The complete-pooling model predicts a significant drop in *within-pair* convergence from the first to second partner, due to the continued influence of the first partner, while the partial-pooling model predicts no drop. We found no evidence of such a drop in the empirical data ($z = -0.66, p = 0.511$), providing further evidence in favor of the full partial-pooling structure.

Discussion

Drawing on general principles of hierarchical Bayesian inference, CHAI suggests that conventions represent the shared structure that agents "abstract away" from partner-specific learning. In this section, we evaluated the extent to which CHAI captured human generalization behavior in a natural-language communication experiment on small networks. Unlike complete-pooling accounts, it 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 are able to converge on shared conventions.

Partner-specificity and generalization presents an even steeper challenge for previous accounts than **P1**. It is not straightforward for previous interactive alignment or reinforcement learning accounts to explain patterns across partner boundaries without being augmented with additional social information. If a particular semantic representation has been primed due to precedent in the preceding dialogue, then the shifting identity of the speaker should not necessarily alter its influence (Brennan & Hanna, 2009; Ferreira, Kleinman, Kraljic, & Siu, 2012; Ostrand & Ferreira, 2019). More sophisticated hierarchical memory retrieval accounts that represent different partners as different *contexts* (e.g Polyn, Norman, & Kahana, 2009; Brown-Schmidt, Yoon, & Ryskin, 2015) may allow priming to be modulated in a partner-specific way, but such an account would presuppose that social information like partner identity is already a salient and relevant feature of the communicative environment. Indeed, an 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 process-level candidate for realizing our hierarchical computational-level model.

A frequent concern in prior work using repeated reference games is that improvements in communication over time are due to generic effects of task familiarity and repetition rather than interactive adaptation to a partner's language use (Hupet & Chantraine, 1992). As they get more practice with the task, speakers may simply get better overall at describing images and listeners may learn how to better identify target images. The effects we observe at partner boundaries show that something is being learned beyond pure familiarity with the task: if speakers and listeners were just learning to better describe and identify targets regardless of who their partner is, we would not expect these reversions. These partner-specificity effects clearly rule out the complete pooling model, but cannot rule out a no-pooling model combined with a practice effect. Under this alternative possibility, partner-specific adaptation would be genuine, but the general decrease in utterance length and increase in accuracy with new partners would be due to practice rather than inductive generalization. Our best current evidence against this practice-based explanation lies in our network convergence results: networks as a whole converge to similar short descriptions across partners, and different networks converge to different descriptions, indicating some gradual degree of transfer across partners. Future work may further address these concerns by including filler trials or by manipulating the length of interaction with each partner.

Our account also predicts that similar inductive learning mechanisms would operate not only across different partners but across different contexts containing different referents. By holding the partner constant across different contexts, rather than holding the context constant across different partners, it would be possible to test the extent to which additional experience along one axis of generalization would affect generalization along the other axis. Finally, one subtler point, which we believe is a rich direction for future research, is how generalization may still depend on the speaker's beliefs about *how partners are sampled*, manifested in their inductive biases at the community-level (Eq. 8). If they believe they are in a tight-knit community where different partners are experts with the domain and have likely interacted with one another before, they may generalize differently than if they believe their community has higher turnover and many novices, brand-new to the task (Isaacs & Clark, 1987).

Phenomenon #3: Conventions are shaped by communicative context

In the previous two sections, we evaluated a model of rapid, partner-specific learning that allows agents to form stable but arbitrary *ad hoc* conventions with partners that gradually generalize to their entire community. The final phenomenon we consider is the way that *ad hoc* conventions are shaped by the communicative needs of the context in which they form. This phenomenon is most immediately motivated by recent findings that more informative or diagnostic words in the local referential context are significantly more likely
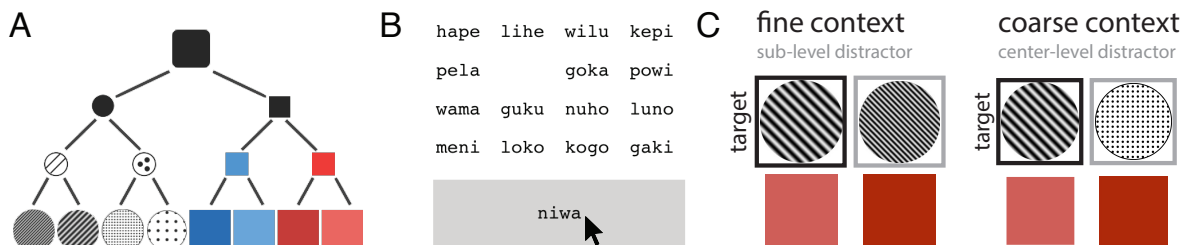
Figure 8

Context-sensitivity experiment. *(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.*

to become conventionalized (Hawkins, Frank, & Goodman, 2020). For example, consider an initial description like "the guy that looks like an ice skater with a leg up in front." A word like "skater," which is distinctive of that single referent, is empirically more likely to persist in the resulting convention than words like "guy" or "leg" which are used in descriptions for multiple referents. Our broader theoretical aim is to suggest that context-sensitivity in the *synchronic* processes at play when individual dyads coordinate on *ad hoc* meanings may help to explain *diachronic* balance of efficiency and expressivity in the long-term evolution of a community's lexicon, as highlighted by functionalist accounts like the Optimal Semantic Expressivity (OSE) hypothesis (Frank, 2017).

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 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[9]. For example, while most English speakers have the general term "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 planted), communicative partners are able to coordinate on efficient lexical conventions for successfully doing so at the relevant level of abstraction (e.g. "the mossy one").

We begin by showing that this form of context-sensitivity naturally emerges from our model, as a downstream consequence of recursive pragmatic reasoning. When a particular partner uses a label to refer to an object in a context, we can infer that they do not believe it ambiguously applies to distractors as well; otherwise, they would have known it would be confusing and chosen a different label. We then empirically evaluate this prediction 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.

---

[9]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 or variability across partners is too high. They must be re-negotiated with subsequent partners on an *ad hoc* basis.

Model predictions: Simulation 3.1

To evaluate the impact of context on convention formation, we require a different task than we used in 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 top-most level, color/texture at the intermediate levels, and frequency/intensity at the finest levels (see Fig. 8A). 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 four squares). We populate the space of possible utterance meanings $P(\phi)$ with four 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"), 1 meaning at the super-ordinate level (e.g. $\phi(u)$ = "square"). We 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. We then populate the utterance space with 8 single-word labels (Fig. 8B).

Another important feature of real environments is that 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 two of the four possible objects 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. 8C).

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 targets sometimes appear in *fine* contexts with subordinate distractors and other times appear in *coarse* contexts without them; the context type is randomized between these two possibilities on each trial. We constructed the trial sequence identically for the three conditions. On each trial, we randomly sampled one of the four possible objects to be the
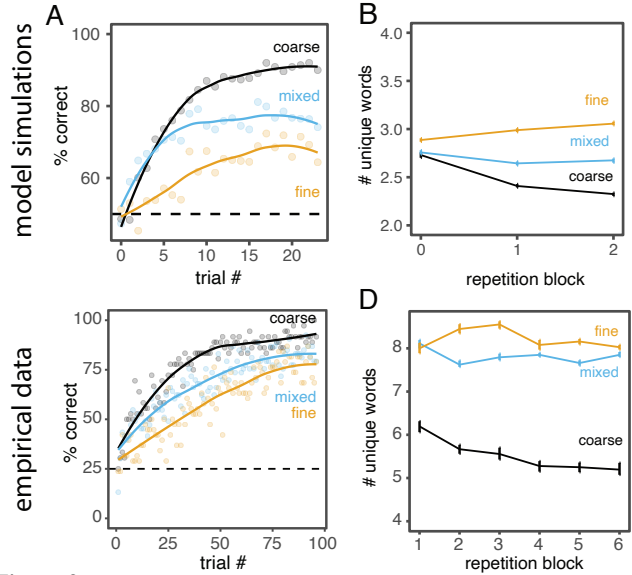

Figure 9

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 stayed roughly constant in the fine condition but decreased over time 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 and error bars are bootstrapped 95% CIs.*

target. Then we sampled a distractor according to the constraints of the context type. As before, the agents swapped roles after each trial. We ran 400 distinct trajectories with parameter settings of $\alpha_L = 8, \alpha_S = 8$ and memory discounting parameter of $\beta = 0.8$ (see Fig. A6 for results at other parameter values).

Partners successfully learn to communicate

First, we compare the model's learning curves across context conditions (Fig. 9A). In a mixed-effects logistic 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.
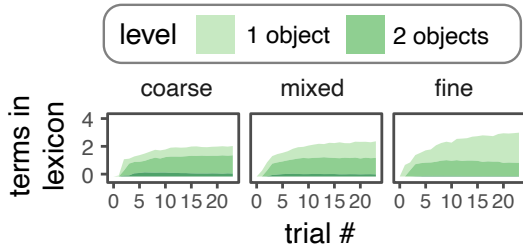
Figure 10

Dynamics of lexical beliefs over time in model simulations. *Regions represent the average number of words at each level of generality in an agent's beliefs about the lexicon. Level of generality is determined by taking the MAP meaning. In the fine and mixed conditions, agents shift toward more subordinate terms.*

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 more 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 dropped over time in the coarse condition, while it stayed roughly constant in in the fine condition, $b = 0.18$, $t = 8.1$, $p < 0.001$, see Fig. 9B.

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. 10, 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. 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. Meanwhile, agents in the fine condition begin to disaggregate these basic-level terms into a greater number 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, multi-player game. 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[10].

Procedure & Stimuli

Participants were paired over the web and placed in a shared environment containing an array of four objects at

---

[10]This experiment was pre-registered at https://osf.io/2hkjc/ with a target sample size of roughly 40 games per condition. We planned to include all participants for our accuracy analyses but then exclude participants who were still below 75% accuracy on the final quarter of the task ($n = 29$ pairs) for our analyses of the lexicon, to ensure post-test measurements could be interpreted as "converging" lexicons (as opposed to pairs who had lost interest or given up). We were later concerned that this exclusion could lead to spurious differences because convergence rates differed across conditions, but no results substantially changed depending on the exclusion criteria. All statistical tests in mixed-effects models reported in this section use degrees of freedom based on the Satterthwaite approximation (Luke, 2017).

a time (Fig. 8A) and a 'chatbox' to choose utterances from a fixed vocabulary by clicking-and-dragging (Fig. 8B). 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. 8B). These arrays were held constant across trials. However, as in our simulations, the set of referents on each trial was manipulated in a between-subjects design to test the context-sensitivity of the resulting conventions. The trial sequence consisted of 6 blocks of 16 trials, for a total of 96 trials. Each of the eight possible objects shown in Fig. 8A appeared as the target exactly twice per block, and was prevented from being shown twice in a row. 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 checking 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'). We conservatively take a participant's final lexicon to be the *intersection* of their word-to-object and object-to-word responses.

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. 9C). 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,
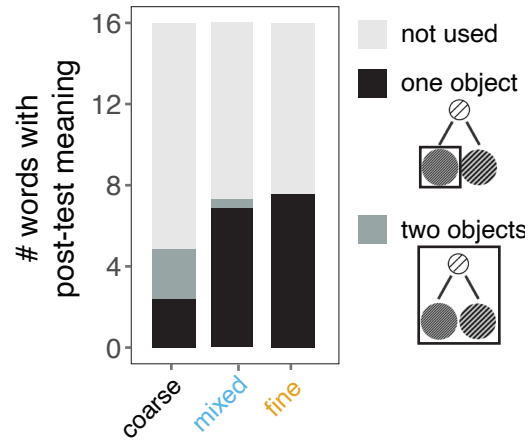

Figure 11

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.*

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 slower than predicted in our simulations. The additional difficulty participants' experienced in the *fine* condition may be due to additional motivational constraints, memory constraints, or other factors not captured in our model.

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. We calculated the same measure of effective vocabulary size that we used in our simulations (Fig. 9D): the number of unique words produced in each repetition block. We then constructed a mixed-effects regression model predicting effective vocabulary size, including fixed effects of condition and six repetition blocks, with random

intercepts and repetition block slopes for each dyad. First, we found an overall main effect of condition, with significantly fewer words used in the coarse condition ($m = 5.5$) than the mixed ($m = 7.9$, $t(95) = 13.2$, $p < 0.001$) or fine ($m = 8.3$, $t(95) = 13.3$, $p < 0.001$) conditions. Consistent with our simulations, we also found a significant interaction between block and condition, with the coarse condition decreasing more over time than mixed ($b = 0.39$, $t(95) = 3.4$, $p < 0.001$) or fine ($b = 0.36$, $t(95) = 2.6$, $p = 0.009$; see Fig. 9D).

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 turned to our post-test survey. We counted the numbers of 'specific' terms (e.g. words that refer to only one object) and '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 mixed ($m = 0.47$, $t(91.8) = 8.8$, $p < 0.001$) or fine ($m = 0.04$, $t(90.2) = 9.2$, $p < 0.001$) 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. A7).

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 are adapted to their immediate needs. In this section, we evaluated predictions about these context effects 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 more sophisticated Discrimination Game (Steels & Belpaeme, 2005; Baronchelli, Gong, Puglisi, & Loreto, 2010), contexts were randomly generated on each trial, but have not been manipulated to assess context-sensitivity of the convention formation process.

In other words, to the extent that context-sensitivity has been addressed by existing models, it has been implicit. Models using simple update rules have accounted for local 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 form of statistical preemption emerges as a natural consequence of normative Gricean principles of pragmatic reasoning rather than as a heuristic (see also Appendix C for similar results using a alternative priors.)

General Discussion

Communication in a variable and non-stationary landscape of meaning creates unique computational challenges. To address these challenges, 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 for prior models of coordination and convention formation: why referring expressions shorten over repeated interactions with the same partner (**P1**), how partner-specific common ground coexists with the emergence of conventions at the population level (**P2**), and how context shapes which conventions emerge (**P3**).

We conclude by raising three broader questions that arise from the perspective of our model, each suggesting pathways for future work: (1) to what extent is *ad hoc* convention formation in adults the same as word learning in children and how is it different? (2) to what extent do the proposed mechanisms depend on the communication modality? and (3) which representations are involved in adaptation at a process-level?

Continuity of language learning across development

CHAI aims to shift the central computational problem of communication from transmission to learning and adapation. Although it is intended as a theory of *adult* communication among mature language users, our emphasis on learning has much in common with theories of language acquisition in development. Could the basic cognitive mechanisms allowing adults to coordinate on conventions be the same as those supporting learning in children? In other words, is it possible that adults never stop learning language and simply develop better-calibrated priors? In this section, we discuss three possible implications of viewing language acquisition in terms of social coordination and convention formation, which may help to further unify models of adult communication with those of language learning (e.g. Xu & Tenenbaum, 2007; Frank et al., 2009; Bohn & Frank, 2019).

First, developmental paradigms have typically focused on variability and generalization across referential contexts (e.g. in cross-situational word learning) rather than variability and generalization across speakers (Siskind, 1996; Regier, 2005; Smith, Suanda, & Yu, 2014; Yurovsky & Frank, 2015). Yet it is increasingly apparent that children are able to track *who* produced the words they are learning and use this information when generalizing. For example, bilingual children learn to expect different languages to be used by different speakers and 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). Children are also sensitive to the reliability of individual speakers. For example, young children may limit the generalizability of observations from speakers who use language in idiosyncratic ways, such as a speaker who calls a ball a "dog" (Koenig & Woodward, 2010; Luchkina, Sobel, & Morgan, 2018), and may even retrospectively update their beliefs about earlier evidence from a speaker after observing such idiosyncrasies (Dautriche et al., 2021). Such 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 equally informative about a community's conventions, CHAI predicts that children should be able to explain away "outliers" without their community-level expectations being disrupted. One 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, CHAI emphasizes the importance of representing lexical *uncertainty* (**C1**), capturing expected variability in the population beyond the point estimates assumed by traditional lexical representations. But how do children calibrate their lexical uncertainty? The number of distinct speakers in a child's environment may play a key role, by analogy to the literature on talker variability (Creel & Bregman, 2011; Clopper & Pisoni, 2004). Exposure to fewer partners may result in weaker or mis-calibrated priors (e.g. Lev-Ari, 2017). 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, reliable community-level conventions may be inferred even with relatively 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 struggle to coordinate *ad hoc* conventions with one another in repeated reference games (Glucksberg, Krauss, & Weisberg, 1966; Krauss & Glucksberg, 1969, 1977; Matthews, Lieven, & Tomasello, 2007). Early explanations appealed to rigidity in the child's perspective that prevented adaptation. Yet subsequent findings that children could not even interpret their own utterances after a delay (Asher & Oden, 1976) suggest that the challenge may instead stem from production quality and the lack of coordinating 'signal'. Children may either be unable to anticipate how much information is required for their partner to discriminate the referent, or struggle to access those more complex formulations. In terms of our model, children's lexical priors may be weaker than adults': without existing conventions for describing the novel objects in their vocabulary, their utterances are dispersed widely over easier-to-access "good-enough" formulations (Goldberg, 2019). Indeed, when children are paired with their caregivers rather than peers, they easily coordinate on new conventions (Leung, Hawkins, & Yurovsky, 2020). Adults helped to interactively scaffold the conventions, both by proactively seeking clarification when in the listener role (e.g. Anderson, Clark, & Mullin, 1994) and by providing more descriptive labels when in the speaker role, which children immediately adopted[11]. From this perspective, *ad hoc* conventions may not be so different from

---

[11] It may be observed that agents in our simulations were still able to quickly coordinate despite being initialized with weak priors, but they had the benefit of using feedback from the referential task, as well as small, shared vocabularies. In the paradigms used by Krauss and Glucksberg (1969), young children did not have access to such information and may have struggled to search their vocabulary for better candidates even if they did, especially under time pressure (e.g. Glucksberg & Krauss, 1967). This kind of accessibility consideration has previously been instantiated in computational models via the cost term $c(u)$, but further work on convention formation in developmental samples may benefit from a more fine-grained, process model of production.

other settings where children look to adults for guidance and rapidly adopt new conventions to talk about new things (e.g. Carey & Bartlett, 1978; Heibeck & Markman, 1987).

The role of communication modality

One of our core claims is that the basic learning mechanisms underlying coordination 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* convention formation dynamics because in every case people will be trying to infer the system of meaning being used by their partners. 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. In fact, there has been significant progress in understanding the dynamics of adaptation during communication in the graphical modality (Garrod et al., 2007; Theisen, Oberlander, & Kirby, 2010; Hawkins, Sano, et al., 2019), the gestural modality (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, Gauvrit, Gibson, & Jamieson, 2019).

CHAI views the similarities and differences between modalities through the lens of the hierarchical *priors* we have built up across interactions with different individuals. For example, in the verbal modality, the tangram shapes from Clark and Wilkes-Gibbs (1986) are highly "innominate" (meaning empirically difficult to name; Hupet, Seron, & Chantraine, 1991; Zettersten & Lupyan, 2020) – most people do not have much experience naming or describing them with words, so relevant 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 therefore suggest that differences in production and comprehension across modalities may be understood by coupling modality-specific priors with modality-generic learning mechanisms.

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 which emphasize current limitations and areas of future work: (1) exactly which representations should be adapted? and (2) what is required to scale models of adaptation to more naturalistic language?

Which representations are adapted?

While our model formulation focused on adaptation at the level of lexical meaning (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 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, expectations about other lower-level features may also be adapted through interaction, such as a partner's word frequencies (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). This level of adaptation may lead some forms to become more accessible or entrenched in memory over time, possibly allowing partner identity to be used as a retrieval cue (e.g. Horton & Gerrig, 2005; Horton, 2007; Horton & Gerrig, 2016; but see Brown-Schmidt & Horton, 2014).

Computational tractability and scalability

While a fully Bayesian formulation elegantly formalizes the computational-level inference problem at the core of the CHAI account, this formulation faces a number of limitations. For one, it is clearly intractable (Van Rooij, 2008; Van Rooij, Blokpoel, Kwisthout, & Wareham, 2019): the posterior update step in Eq. 5 grows increasingly intensive as the space of possible utterances and meanings grows (Woensdregt, Spike, de Haan, van Rooij, & Blokpoel, 2021). The intractability problem also raises a scalability problem: does CHAI provide any guidance toward building artificial agents that are actually able to adapt to human partners as humans do with one another? Through this applied lens, a number of recent efforts have focused on developing algorithms for state-of-the-art neural networks that tractably scale 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).

For example, building on recent formal connections between hierarchical Bayes and gradient-based meta-learning approaches in machine learning (Grant, Finn, Levine, Darrell, & Griffiths, 2018), the algorithm proposed by Hawkins, Kwon, et al. (2020) (1) relaxes the full community-level prior over Θ to a point estimate and (2) replaces the difficult integral in the posterior update with a fixed number of (regularized) gradient update steps. Another recent proposal builds on connections to classical exemplar-based algorithms (Nosofsky, 1984): an agent's lexical expectations at time t may be determined via weighted similarity to memory traces of lexical items used by different partners in the past (Shi, Griffiths, Feldman, & Sanborn, 2010), where similarity is computed by a neural network. While such algorithms cannot fix the intractability of the Bayesian formulation (Kwisthout, Wareham, & van Rooij, 2011), and their precise correspondence to constraints on the computational-level theory remain unexplored, they nonetheless provide promising algorithmic instantiations of the CHAI account. When lexical meaning is represented by the parameters of a neural network, conventions can be interpreted as (meta-)learned *initializations* used for new partners and coordination is partner-specific *fine-tuning* or *domain adaptation* of vector representations.

Neural network instantiations also provide a possible pathway toward addressing the lack of incrementality in the fully Bayesian formulation. As more scalable implementations of pragmatic reasoning have proliferated in machine learning (Vogel, Bodoia, Potts, & Jurafsky, 2013; Andreas & Klein, 2016; Monroe et al., 2017; Shen, Fried, Andreas, & Klein, 2019; Takmaz, Giulianelli, Pezzelle, Sinclair, & Fernández, 2020) it has been natural to use incremental architectures (Augurzky, Franke, & Ulrich, 2019; Cohn-Gordon, Goodman, & Potts, 2018, 2019; Waldon & Degen, 2021). However, there remain a number of limitations to address in future work, including how to incorporate incremental *feedback* into lexical updates (e.g. backchannel responses or interruptions), how to define a more satisfying notion of *compositional semantics* for incrementally constructed utterances, and how to maintain representations of partner-specific parameters alongside community-wide parameters in memory.

Conclusion. How do we manage to understand one another? We have argued that successful communication depends not just on transmission but on continual learning across multiple timescales. We must coordinate on meaning through common ground with individual partners but also abstract these experiences away to represent stable *conventions* and norms that generalize across our communities. Like other socially-grounded knowledge, language is not a rigid dictionary that we acquire at an early age and deploy mechanically for the rest of our lives. Nor do languages only change over the slow time-scales of inter-generational drift. Language is a means for communication – a shared interface between minds – and as new *ad hoc* concepts arise, new *ad hoc* conventions must be formed to solve the new coordination problems they pose. In other words, we are constantly learning language. Not just one language, but a family of related languages, across interactions with each 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)

Acknowledgments

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.

Achlioptas, P., Fan, J., Hawkins, R., Goodman, N., & Guibas, L. J. (2019). ShapeGlot: Learning language for shape differentiation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 8938–8947).

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.* New York, NY: Oxford University Press.

Andreas, J., & Klein, D. (2016). Reasoning about pragmatics with neural listeners and speakers. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 1173–1182).

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

Arkel, J. v., Woensdregt, M., Dingemanse, M., & Blokpoel, M. (2020). A simple repair mechanism can alleviate computational demands of pragmatic reasoning: Simulations and complexity analysis. In *Proceedings of the 24th conference on computational natural language learning* (pp. 177–194).

Armstrong, J. (2016a). Coordination, triangulation, and language use. *Inquiry*, *59*(1), 80–112.

Armstrong, J. (2016b). The problem of lexical innovation. *Linguistics and Philosophy*, *39*(2), 87–118.

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*. Abingdon: Routledge.

Augurzky, P., Franke, M., & Ulrich, R. (2019). Gricean expectations in online sentence comprehension: An erp study on the processing of scalar inferences. *Cognitive Science*, *43*(8), e12776.

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.

Barsalou, L. W. (1983). Ad hoc categories. *Memory & Cognition*, *11*(3), 211–227.

Bergen, L., Goodman, N. D., & Levy, R. (2012). That's what

she (could have) said: How alternative utterances affect language use. In *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (p. 120-125).

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*. Cambridge, MA: 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).

Brochhagen, T. (2020). Signalling under uncertainty: Interpretative alignment without a common prior. *The British Journal for the Philosophy of Science*, *71*(2), 471–496.

Brochhagen, T. (2021). Brief at the risk of being misunderstood: Consolidating population-and individual-level tendencies. *Computational Brain & Behavior*, *4*, 305-317.

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

Brown-Schmidt, S., & Horton, W. S. (2014). The influence of partner-specific memory associations on picture naming: A failure to replicate Horton (2007). *PloS one*, *9*(10), e109035.

Brown-Schmidt, S., Yoon, S. O., & Ryskin, R. A. (2015). People as contexts in conversation. *Psychology of Learning and Motivation*, *62*, 59–99.

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*, *195*, 603--621.

Carey, S., & Bartlett, E. (1978). Acquiring a single new word. *Papers and Reports on Child Language Development*, *15*, 17-29.

Carr, J. W., Smith, K., Culbertson, J., & Kirby, S. (2020). Simplicity and informativeness in semantic category systems. *Cognition*, *202*, 104289.

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*. New York, NY: Cambridge University Press.

Clark, E. V., & MacWhinney, B. (1987). The principle of contrast: A constraint on language acquisition. *Mechanisms of language acquisition*, 1–33.

Clark, H. H. (1996). *Using language*. New York, NY: Cambridge University Press.

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

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

Clark, H. H., & Marshall, C. (1981). Definite reference and mutual knowledge. In A. Joshi, B. Webber, & I. Sag (Eds.), *Elements of discourse understanding* (p. 10–63). Cambridge Unviersity Press.

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

Cohn-Gordon, R., Goodman, N., & Potts, C. (2018). Pragmatically informative image captioning with character-level inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)* (pp. 439–443).

Cohn-Gordon, R., Goodman, N., & Potts, C. (2019). An incremental iterated response model of pragmatics. In *Proceedings of the Society for Computation in Linguistics (SCiL)* (pp. 81–90).

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.

Dingemanse, M., Roberts, S. G., Baranova, J., Blythe, J., Drew, P., Floyd, S., . . . Enfield, N. J. (2015). Universal principles in the repair of communication problems. *PloS one*, *10*(9), e0136100.

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

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.

Effenberger, A., Yan, E., Singh, R., Suhr, A., & Artzi, Y. (2021). Analysis of language change in collaborative instruction following. *arXiv preprint arXiv:2109.04452*.

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.

Ferreira, V. S., Kleinman, D., Kraljic, T., & Siu, Y. (2012). Do priming effects in dialogue reflect partner-or task-based expectations? *Psychonomic Bulletin & Review*, *19*(2), 309–316.

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 coordination. *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.

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

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

Gerrig, R. J., & Gibbs Jr, R. W. (1988). Beyond the lexicon: Creativity in language production. *Metaphor and Symbol*, *3*(3), 1–19.

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

Gershman, S. J., & Cikara, M. (2020). Social-structure learning. *Current Directions in Psychological Science*, *29*(5), 460–466.

Gershman, S. J., & Niv, Y. (2015). Novelty and inductive generalization in human reinforcement learning. *Topics in cognitive science*, *7*(3), 391–415.

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. (1967). What do people say after they have learned how to talk? studies of the development of referential communication. *Merrill-Palmer Quarterly of Behavior and Development*, *13*(4), 309–316.

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.

Grosz, B. (1974). The structure of task oriented dialogs. In *IEEE Symposium on Speech Recognition* (Vol. 10).

Grosz, B., & Sidner, C. L. (1986). Attention, intentions, and the structure of discourse. *Computational Linguistics*, *12*(3), 175-204.

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 soci-*

ety.

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., Liu, I., Goldberg, A. E., & Griffiths, T. L. (2021). Respect the code: Speakers expect novel conventions to generalize within but not across social group boundaries. In *Proceedings of the 43rd annual meeting of the cognitive science society*.

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). Oxford: Oxford University Press.

Heibeck, T. H., & Markman, E. M. (1987). Word learning in children: An examination of fast mapping. *Child development*, 1021–1034.

Horn, L. (1984). Toward a new taxonomy for pragmatic inference: Q-based and R-based implicature. *Meaning, form, and use in context: Linguistic applications*, *11*, 42.

Horton, W. S. (2007). The influence of partner-specific memory associations on language production: Evidence from picture naming. *Language and cognitive processes*, *22*(7), 1114–1139.

Horton, W. S., & Gerrig, R. J. (2005). The impact of memory demands on audience design during language production. *Cognition*, *96*(2), 127–142.

Horton, W. S., & Gerrig, R. J. (2016). Revisiting the memory-based processing approach to common ground. *Topics in Cognitive Science*, *8*, 780–795.

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 de-

vice. *Lingua*, *77*(2), 187–222.

Ibarra, A., & Tanenhaus, M. K. (2016). The flexibility of conceptual pacts: referring expressions dynamically shift to accommodate new conceptualizations. *Frontiers in psychology*, *7*.

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

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.

Kidd, E., Donnelly, S., & Christiansen, M. H. (2018). Individual differences in language acquisition and processing. *Trends in cognitive sciences*, *22*(2), 154–169.

Kinzler, K. D. (2021). Language as a social cue. *Annual Review of Psychology*, *72*, 241–264.

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.

Kwisthout, J., Wareham, T., & van Rooij, I. (2011). Bayesian intractability is not an ailment that approximation can cure. *Cognitive Science*, *35*(5), 779–784.

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

Lakoff, G. (1975). Hedges: A study in meaning criteria and the logic of fuzzy concepts. In *Contemporary research in philosophical logic and linguistic semantics* (pp. 221–271). Springer.

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:2101.01951*.

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.

Levinson, S. C., Stephen, C., & Levinson, S. C. (2000). *Presumptive meanings: The theory of generalized conversational implicature.* MIT press.

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.

Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis.* New York, NY: Wiley.

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.

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.

McCarthy, W. P., Hawkins, R., Wang, H., Holdaway, C., & Fan, J. E. (2021). Learning to communicate about shared procedural abstractions. In *Proceedings of the 43rd annual meeting of the cognitive science society* (p. 77-83).

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).

Misyak, J. B., Melkonyan, T., Zeitoun, H., & Chater, N. (2014). Unwritten rules: virtual bargaining underpins social interaction, culture, and society. *Trends in Cognitive Sciences*, *18*(10), 512–519.

Monroe, W., Hawkins, R. D., Goodman, N. D., & Potts, C. (2017). Colors in context: A pragmatic neural model for grounded language understanding. *Transactions of the Association for Computational Linguistics*, *5*, 325–338.

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.

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, memory, and cognition*, *10*(1), 104.

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.*

Ostrand, R., & Ferreira, V. S. (2019). Repeat after us: Syntactic alignment is not partner-specific. *Journal of memory and language*, *108*, 104037.

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.

Piantadosi, S. T., Tily, H., & Gibson, E. (2012). The communicative function of ambiguity in language. *Cognition*, *122*(3), 280–291.

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., Lassiter, D., Levy, R., & Frank, M. C. (2016). Embedded implicatures as pragmatic inferences under compositional lexical uncertainty. *Journal of Semantics*, *33*(4), 755–802.

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 rea-

soning. 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.

Rosenberg, S., & Cohen, B. D. (1966). Referential processes of speakers and listeners. *Psychological Review*, *73*(3), 208–231.

Schegloff, E. A., Jefferson, G., & Sacks, H. (1977). The preference for self-correction in the organization of repair in conversation. *Language*, 361–382.

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 introduction 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*.

Shen, S., Fried, D., Andreas, J., & Klein, D. (2019). Pragmatically informative text generation. *arXiv preprint arXiv:1904.01301*.

Shi, L., Griffiths, T. L., Feldman, N. H., & Sanborn, A. N. (2010). Exemplar models as a mechanism for performing bayesian inference. *Psychonomic bulletin & review*, *17*(4), 443–464.

Shoham, Y., & Tennenholtz, M. (1997). On the emergence of social conventions: modeling, analysis, and simulations. *Artificial Intelligence*, *94*(1-2), 139–166.

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.

Takmaz, E., Giulianelli, M., Pezzelle, S., Sinclair, A., & Fernández, R. (2020). Refer, reuse, reduce: Generating subsequent references in visual and conversational contexts. *arXiv preprint arXiv:2011.04554*.

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 de Braak, L. D., Dingemanse, M., Toni, I., van Rooij, I., & Blokpoel, M. (2021). Computational challenges in

explaining communication: How deep the rabbit hole goes. In *Proceedings of the 43rd annual meeting of the cognitive science society* (p. 528-534).

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 Rooij, I. (2008). The tractable cognition thesis. *Cognitive science*, *32*(6), 939–984.

Van Rooij, I., Blokpoel, M., Kwisthout, J., & Wareham, T. (2019). *Cognition and intractability: A guide to classical and parameterized complexity analysis*. Cambridge University Press.

Vélez, N., & Gweon, H. (2021). Learning from other minds: An optimistic critique of reinforcement learning models of social learning. *Current Opinion in Behavioral Sciences*, *38*, 110–115.

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).

Vogel, A., Bodoia, M., Potts, C., & Jurafsky, D. (2013). Emergence of gricean maxims from multi-agent decision theory. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies* (pp. 1072–1081).

Waldon, B., & Degen, J. (2021). Modeling cross-linguistic production of referring expressions. *Proceedings of the Society for Computation in Linguistics*, *4*(1), 206–215.

Wang, X., & Bi, Y. (2021). Idiosyncratic tower of babel: Individual differences in word-meaning representation increase as word abstractness increases. *Psychological science*.

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.

Woensdregt, M. S., Spike, M., de Haan, W. T., R and, van Rooij, I., & Blokpoel, M. (2021). Why is scaling up models of language evolution hard?.

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*.

Zettersten, M., & Lupyan, G. (2020). Finding categories through words: More nameable features improve category learning. *Cognition*, *196*, 104135.

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.

Action-oriented vs. belief-oriented listeners

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).

Placement of uncertainty

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.

	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.

Handling degenerate lexicons

Finally, 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. 3 is zero, and the literal listener distribution is not well-defined. A similar problem may arise for the S_1 distribution.

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 will assign negligible listener probability to objects where u is false, but ensures that the normalization constant is non-zero (and specifically, that the distribution is uniform) 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 later iterations for some parameter values, 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. 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.

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, i.e. a 'failure to refer,' and effectively prevents L_0 from assigning belief to a referent for which the utterance is literally false. Note that this case is distinct from the case of a *contradiction*, which arises when defining the meaning of multi-word utterances in Section **P1**.

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.

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

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

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. 4). 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

$$U_L(o; u, \phi_k) = \log S_1(u|o, \phi_k)$$
$$U_S(u; o, \phi_k) = (1 - w_C) \log L_0(o|u, \phi_k) - w_C \cdot c(u)$$

then the expectation is taken as follow:

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

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.

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

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..

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

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.

Inference details

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. 3), 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. 5), 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. 4 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.

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. $w_1 = 'blue\,square'$, 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:

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

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:

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

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 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).

Appendix C: Alternative lexical priors

A variety of priors have been proposed for probabilistic models of language learning and convention formation, which build in stronger or weaker assumptions about the structure of the lexicon. We used a simplicity prior over lexicons that strictly partition the set of referents, ensuring that every object must be in the extension of exactly one word (e.g. Carr et al., 2020). To explore the robustness of our results in **P3** to alternative choices of lexical priors, we also considered a weaker prior where the space of possible lexicons allows any denotation to be assigned to any word, including highly redundant and overlapping lexicons (where every object is in the extension of every word), and highly degenerate lexicons (where every word has a completely empty extension). Given this more unconstrained space of lexicons, the simplest way to penalize complexity is to define the size of the lexicon $|\phi|$ as the *total extension size*, i.e. the summed extensions of all terms. In this case, favoring simpler word meanings also necessarily favors smaller lexicons.

To illustrate this property of the weaker prior, consider a reference game with a fixed set of two referents: a blue square and a red square. Then the meaning of a given utterance u_i can either have extension size two (applying both objects, effectively meaning 'square'), size one (applying only to 'blue square' or only to 'red square'), or size zero (applying to neither of the objects). Now suppose there are 16 possible utterances. Then the lexicon with highest prior probability is the one where every utterance has an empty extension ($|\phi| = 16 \cdot 0 = 0$), which is also the smallest lexicon (an effective vocabulary size of 0). The lexicon with lowest prior probability is the one where every utterance has the maximal extension ($|\phi| = 16 \cdot 2 = 32$), which is also the largest lexicon (an effective vocabulary size of 16). Removing a single word from this maximal lexicon would reduce the size of the lexicon (e.g. 15 words instead of 16) and also reduce the

total size of the words' extensions (e.g. taking a word with the maximal extension of size two and replacing it with the minimal extension of size zero: $|\phi| = 15\cdot2+1\cdot0 = 30$), which would slightly increase its prior probability. Of course, this scheme also makes it possible to have a lexicon with smaller word meanings but not a smaller lexicon (e.g. for a lexicon where all 16 words have an extension size of 1, we have $|\phi| = 16 \cdot 1 = 16$, but the effective vocabulary size is still 16). So this prior straightforwardly encodes a preference for lexicons with fewer words, but gives partial credit when the words have 'simpler' meanings, breaking ties between lexicons with the same number of words. Simulation results using this prior are shown in Fig. A6B.

Note that other choices are possible, such as those only enforcing mutual exclusivity, or only the principle of contrast (Clark & MacWhinney, 1987). In the context of **P3**, we expect these choices to primarily affect choices on early trials. Some priors may predict an *agglomerative* form of learning where all conditions begin using fine-grained language and then the *coarse* condition gradually collapses down to a more minimal lexicon, while others predict a *divisive* form where all conditions begin using coarse-grained language and then the *fine* condition gradually introduces words with more refined meanings. For example, we one qualitative feature of our empirical results in **P3** (see Fig. 9D) is that participants apparently begin using more unique terms at the outset, and remain constant at that large number of unique terms in the *fine* and *mixed* conditions but gradually whittle away their vocabulary size in the *coarse* condition. Because we noticed that neither of the priors considered in Fig. A6 displayed this

pattern, we considered a third possible prior. This prior enforced full coverage over all meanings (i.e. disallowed 'degenerate' lexicons where some objects are not in the extension of any words at all), unlike the unconstrained prior, but otherwise allowed redundancy (i.e. some objects were in the extension of multiple words), unlike the partition-based prior. This prior gave rise to a qualitatively more similar pattern of lexical convergence (see Fig. A1).
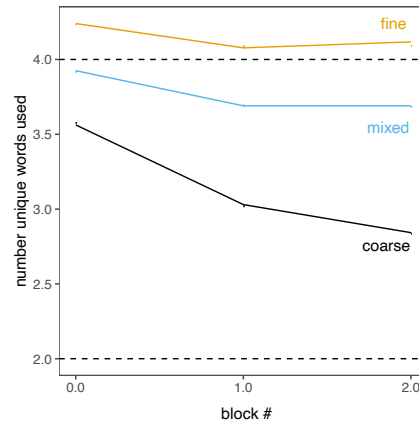

Figure A1

*Simulation results for **P3** using a full-coverage lexical prior that disallows degenerate lexicons ($\alpha_S = \alpha_L = 6, \beta = 0.6$)*
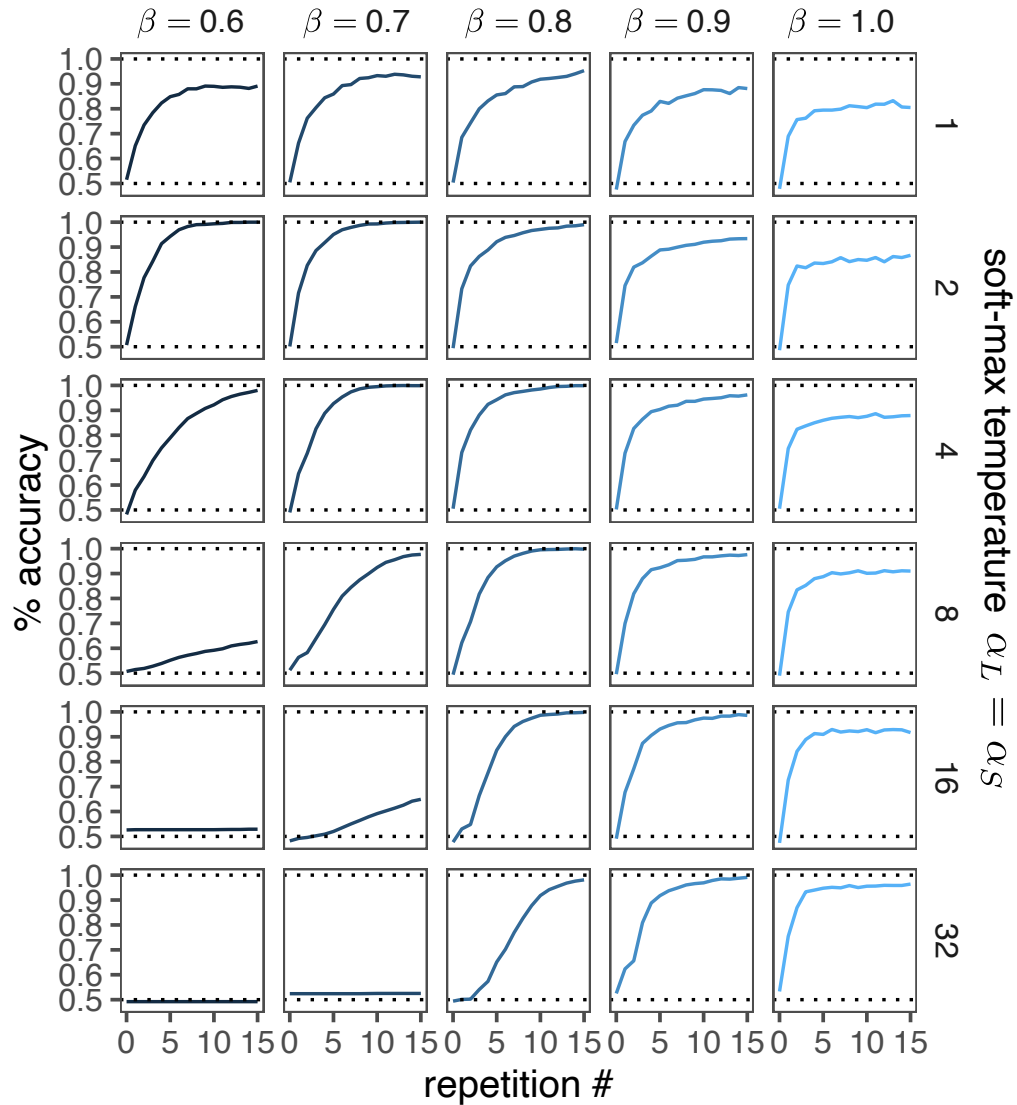
Figure A2

Coordination success (simulation 1.1) across a range of parameter values. Columns represent memory discount parameter β, and rows represent the agents' soft-max optimality parameters, 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.
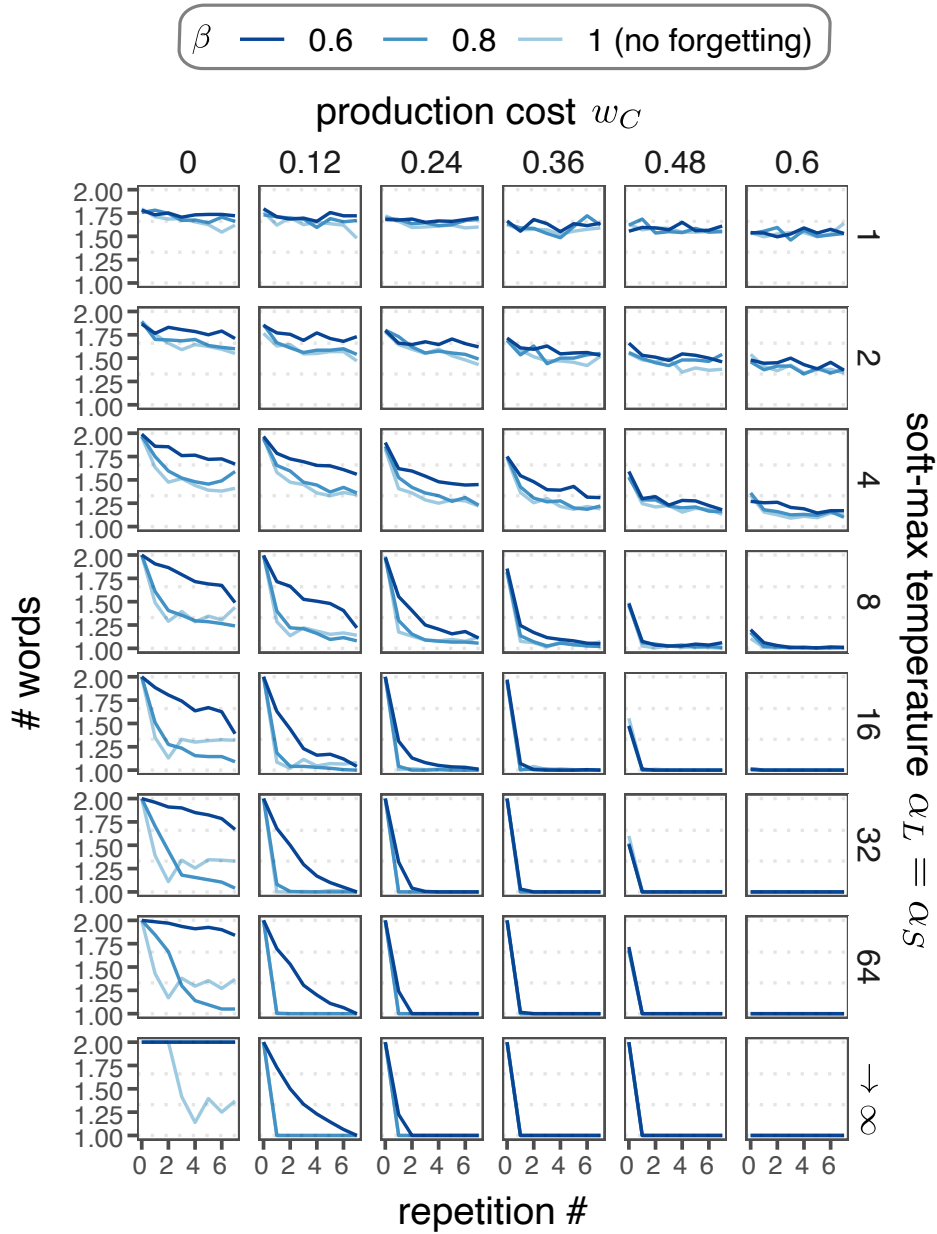
Figure A3

Speaker efficiency (simulation 1.2) across a range of parameter values representing different weights on informativity and cost. Rows represent agent soft-max optimality $\alpha_S = \alpha_L$, 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 α_S, α_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).
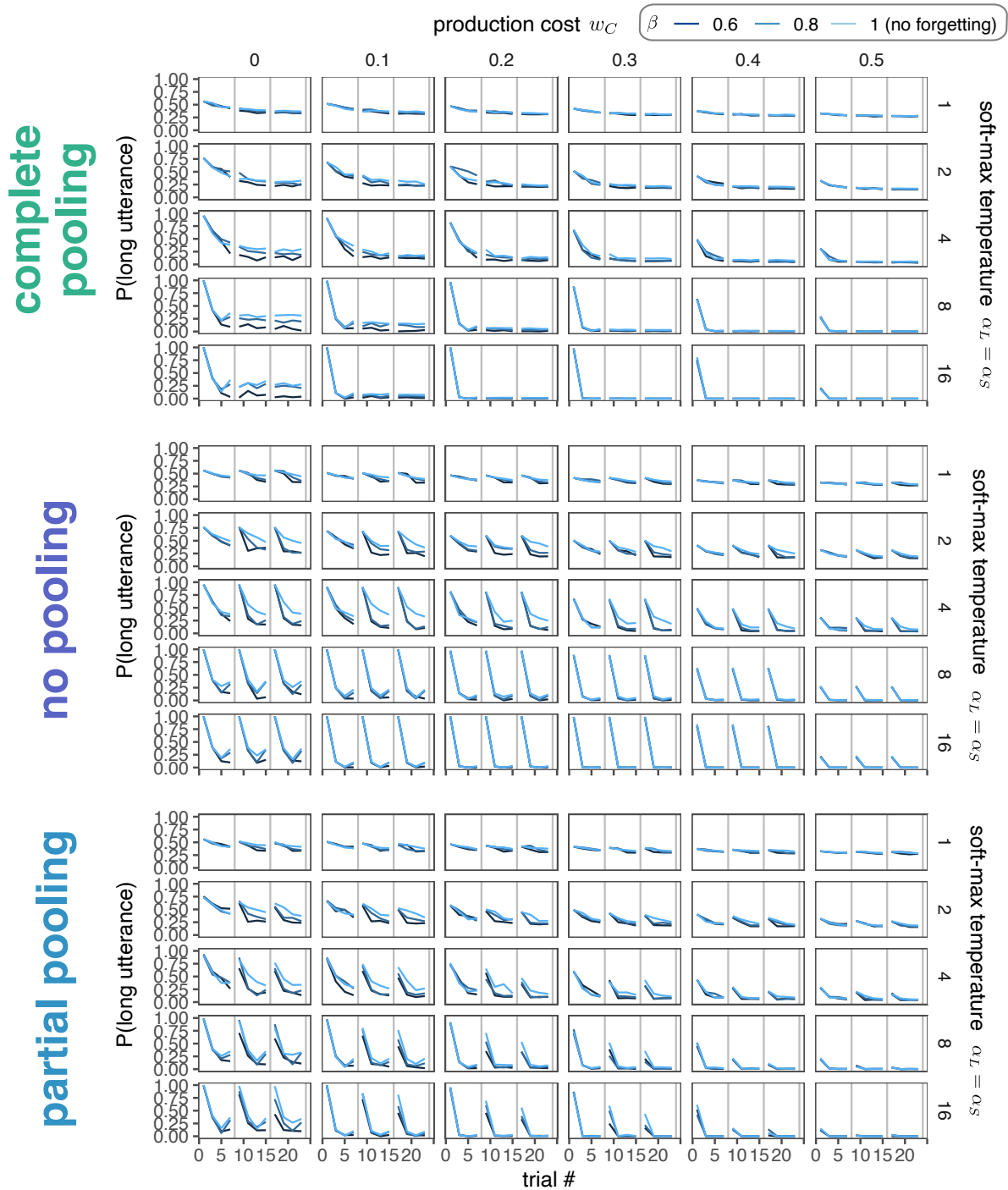
Figure A4

*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 soft-max optimality $\alpha_S = \alpha_L$, columns represent cost weight w_c, and colors represent memory discount parameter β.*
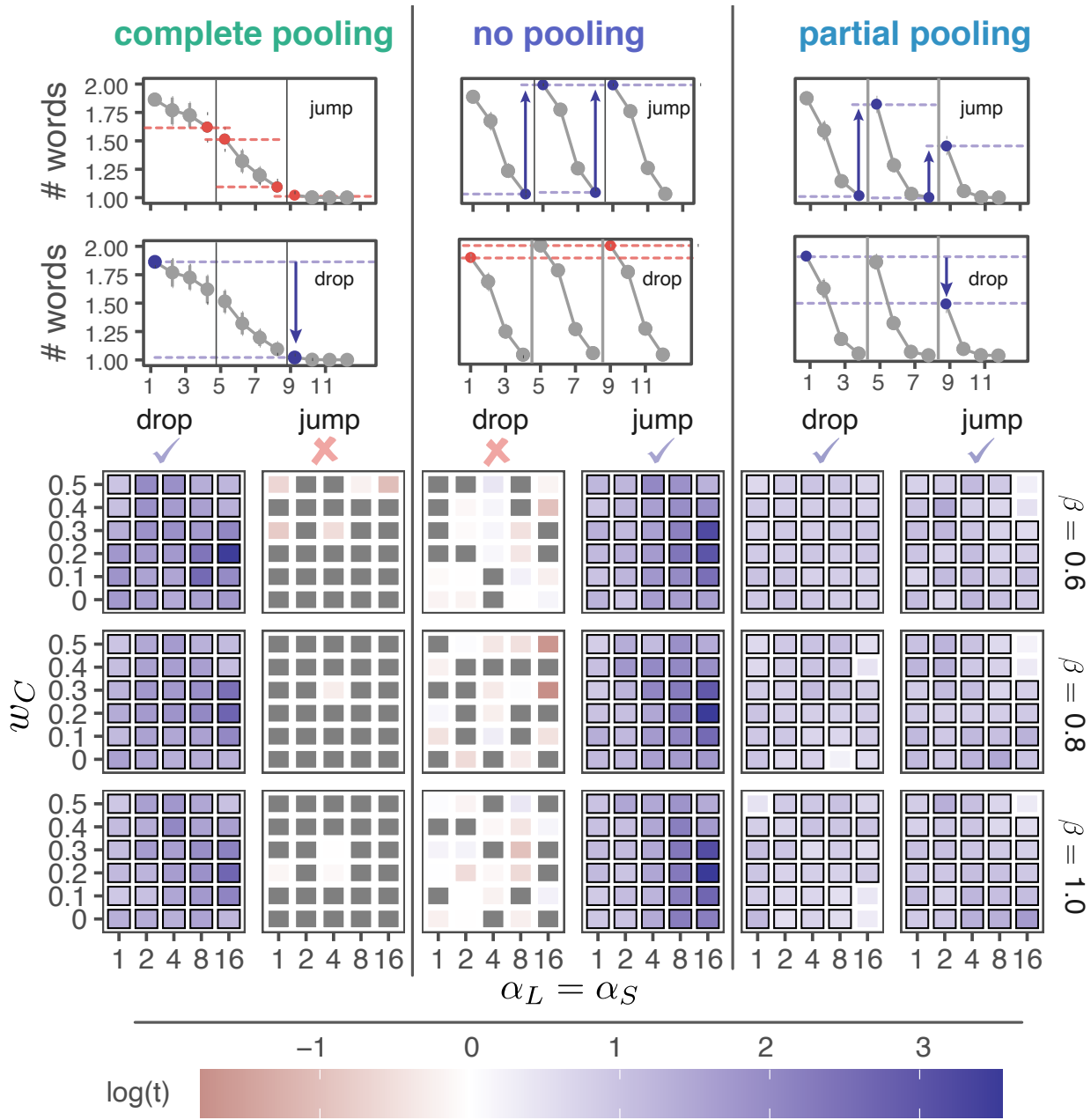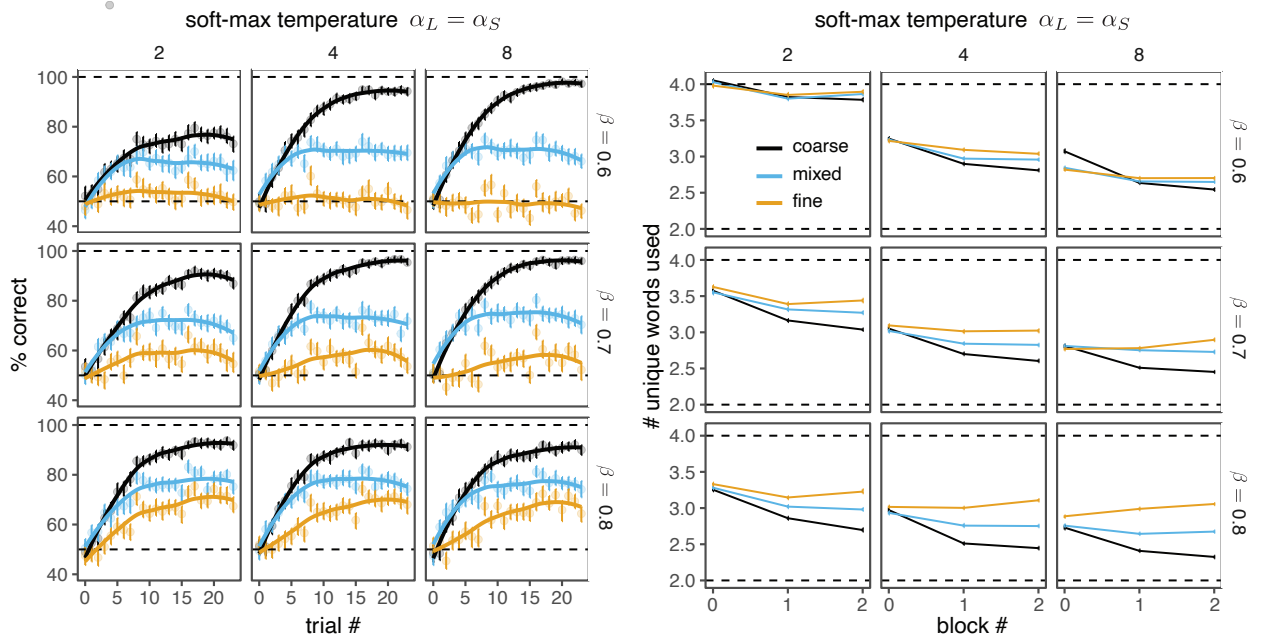
Figure A5

*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.*

A partition-based simplicity prior
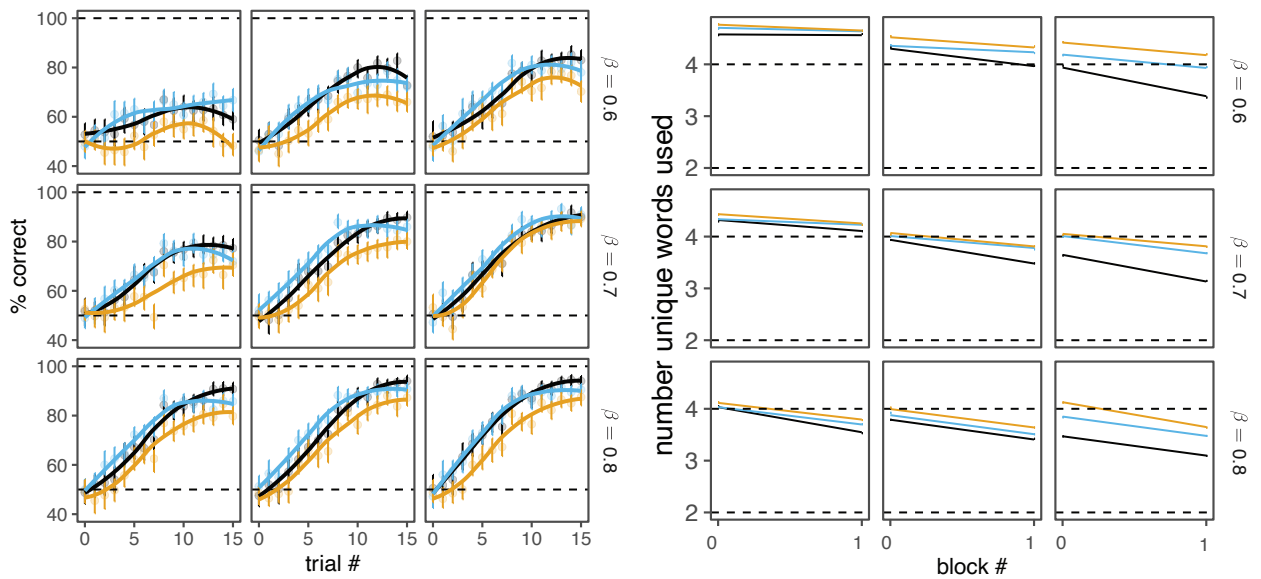

B unconstrained simplicity prior

Figure A6

*Simulation results for **P3** using different values of α and β for (A) partition-based simplicity prior, and (B) alternative unconstrained simplicity prior. Simulations were run for fewer trials in (B). Overall, we observe similar qualitative predictions for the difference between the coarse and fine condition, although the mixture condition is more sensitive to parameters and priors.*
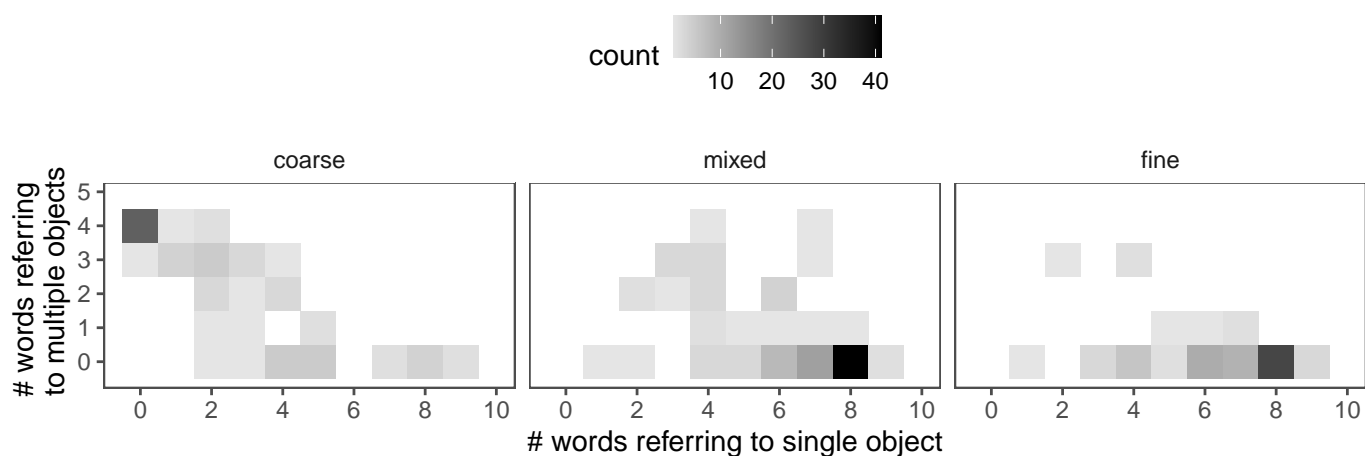
Figure A7

*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.*