Generalizing meanings from partners to populations: Hierarchical inference supports convention formation on networks

Anonymous CogSci submission

Abstract

A key property of linguistic conventions is that they hold over an entire community of speakers, allowing us to communicate efficiently even with people we have never met before. At the same time, much of our language use is partner-specific: we know that words may be understood differently by different people based on local common ground. This poses a challenge for accounts of convention formation. Exactly how do agents make the inferential leap to community-wide expectations while maintaining partner-specific knowledge? We propose a hierarchical Bayesian model of convention to explain how speakers and listeners abstract away meanings that seem to be shared across partners. To evalute our model's predictions, we conducted an experiment where participants played an extended natural-language communication game with different partners in a small community. We examine several measures of generalization across partners, and find key signatures of local adaptation as well as collective convergence. These results suggest that local partner-specific learning is not only compatible with global convention formation but may facilitate it when coupled with a powerful hierachical inductive mechanism.

Keywords: convention; generalization;

To communicate successfully, speakers and listeners must share a common system of semantic meaning in the language they are using. These meanings are *conventional* in the sense that they are sustained by the expectations each person has about others (Bicchieri, 2006; Lewis, 1969). A key property of linguistic conventions is that they hold over an entire community of speakers, allowing us to communicate efficiently even with people we've never met before. But exactly how do we make the inferential leap to community-wide expectations from our experiences with specific partners? Grounding collective convention formation in individual cognition 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 (Baronchelli, 2018; Barr, 2004; Steels, 1995; 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 & 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 even 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). Extensive evidence from psycholinguistics has demonstrated the *partner-specificity* of our language use (Clark, 1996). Because meaning is grounded in the evolving 'common ground' shared with each partner, meanings established over a history of interaction with one partner are not necessarily transfered to other partners (Metzing & Brennan, 2003; Wilkes-Gibbs & Clark, 1992). 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. The problem with no-pooling, of course, is that agents are forced to start from scratch with each partner. Community-level expectations never get off the ground.

What theory of generalization, then, can explain partner-specific meaning but also allow conventions to spread through communities? We propose a *partial-pooling* account that offers a compromise between these extremes. Unlike complete-pooling and no-pooling models, we propose that beliefs about meaning 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 (see also Kleinschmidt & Jaeger, 2015; Tenenbaum, Kemp, Griffiths, & Goodman, 2011).

We begin by formalizing this account in a probabilistic model of communication and presenting several simulations of listener and speaker behavior within and across partners. Next, we test the qualitative predictions of this model in a behavioral experiment. Participants were paired for a series of extended reference games with each neighbor in small networks. Our results showed signatures of *ad hoc* convention formation within dyads, but also gradual generalization of these local pacts across subsequent partners as the network converged. Taken together, these results suggest that local partner-specific learning is not only compatible with global convention formation but may facilitate it when coupled with a powerful hierachical inductive mechanism.

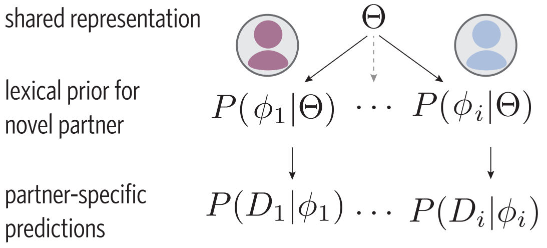
shared representation
Θ
lexical prior for novel partner
$P(\phi_1|\Theta) \cdots P(\phi_i|\Theta)$
partner-specific predictions
$P(D_1|\phi_1) \cdots P(D_i|\phi_i)$

Figure 1: Schematic of hierachical Bayesian model.

A hierarchical Bayesian model of convention

In this section, we provide an explicit computational account of the cognitive mechanisms supporting the balance between community-level stability and partner-specific flexibility. Specifically, we show how the dyadic convention formation model of Hawkins, Frank, & Goodman (2017) can be extended with a principled mechanism for generalization across multiple partners. This model begins with the idea that knowledge about meanings can be represented probabilistically: agents have uncertainty about what lexical meaning their current partner is using (Bergen, Levy, & Goodman, 2016). In our hierarchical model, this lexical 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 possible partners. Θ then parameterizes the agent's *partner-specific* uncertainty $P(\phi_k|\Theta)$, where ϕ_k represents the specific system of meanings used by partner k (see Fig. 1). Given observations D_k from repeated communicative interactions with k, the agent updates their beliefs about the latent system of meaning using Bayes rule:

$$\begin{aligned} P(\phi_k, \Theta | D_k) &\propto P(D_k|\phi_k, \Theta)P(\phi_k, \Theta) \\ &= P(D_k|\phi_k)P(\phi_k|\Theta)P(\Theta) \end{aligned} \tag{1}$$

\aeg{The figure is explained below, but maybe make the individual-images be made slightly smaller so that it's clear they belong to the second row? Also, the 2nd row label on the left also seems to imply any novel partner, but then it's odd to assign a random partner two distinct colors. Isn't it "lexical prior for individuals?"? This joint inference decomposes the problem of partner-specific learning into two terms, a prior term $P(\phi_k|\Theta)P(\Theta)$ and a likelihood term $P(D_k|\phi_k)$. The prior captures the idea that different partners share some aspects of meaning in common. In the absence of strong information about partner-specific language use departing from this common structure, the agent ought to be regularized toward generalizable knowledge of their community's conventions (Davidson, 1986). The likelihood represents predictions about how a partner using a particular system of meaning will use language.

This joint posterior over meanings has two consequences for convention formation. First, it allows agents to maintain partner-specific expectations ϕ_k by marginalizing over community-level uncertainty:

$$P(\phi_k|D_k) = \int_\Theta P(D_k|\phi_k)P(\phi_k|\Theta)P(\Theta)d\Theta \tag{2}$$

Second, the hierarchical structure provides an inductive pathway for data to inform beliefs about community-wide conventions. Agents update their beliefs about Θ by marginalizing over data accumulated from different partners:

$$P(\Theta|D) = P(\Theta)\int_\phi P(D_k|\phi_k)P(\phi_k|\Theta)d\phi \tag{3}$$

where $D = \bigcup_{k=1}^N D_k$, $\phi = \phi_1 \times \cdots \times \phi_N$, and N is the number of partners previous encountered.

After multiple partners are inferred to have a similar system of meaning, beliefs about Θ shift to represent this abstracted knowledge: it becomes more likely that a novel partner will share it as well. This transfer is sometimes referred to as a "sharing of strength" or "partial pooling" (Gelman & Hill, 2006) because pooled data is smoothly integrated with domain-specific detail depending on the data available.

Model simulations

We investigate the qualitative predictions of this model under three increasingly complex scenarios. In all of these scenarios, speaker and listener agents play a reference game with a set of two objects $\{o_1, o_2\}$. On each trial, one of these objects is designated for the speaker as the *target*. They must select from a set of utterance $\{u_0, \ldots, u_j\}$ to convey the identity of the target to the listener. Upon hearing this utterance, the listener selects which of the objects they believe to be the target and then receives feedback about the true target. The resulting data D_k from an interaction with partner k thus consists of utterance-object pairs $\{(u,o)_t\}$ for each trial t, as well as information about the context of objects.

Given this reference game setting, we can now explicitly specify the likelihood and prior terms. We consider a likelihood given by the Rational Speech Act (RSA) framework, which formalizes the Gricean assumption of cooperativity (Franke & Jäger, 2016; Goodman & Frank, 2016). A pragmatic speaker S_1 attempts to trade off informativity against the cost of producing an utterance, while a pragmatic listener L_1 inverts their model of the speaker to infer the intended target. The chain of recursive social reasoning grounds out in a *literal listener* L_0, who identifies an intended meaning using their knowledge of lexical items $\mathcal{L}_{\phi_k}$. This model can be formally specified as follows:

$$\begin{aligned} L_0(o|u,\phi_k) &\propto \exp\{\mathcal{L}_{\phi_k}(u,o)\} \\ S_1(u|o,\phi_k) &\propto \exp\{w_I \cdot \log L_0(o|u,\phi_k) - w_C \cdot \text{cost}(u)\} \\ L_1(o|u,\phi_k) &\propto S_1(u|o,\phi_k)P(o) \end{aligned}$$

where w_I and w_C are free parameters controlling the relative weights on the informativity and parsimony, respectively[1].

[1] Throughout our simulations, we set $w_I = 11$, $w_C = 7$. A grid search over parameter space revealed different behaviors in different regimes (e.g. when cost was weighted too strongly, multi-word utterances were not produced); we leave broader exploration of this space to future work.

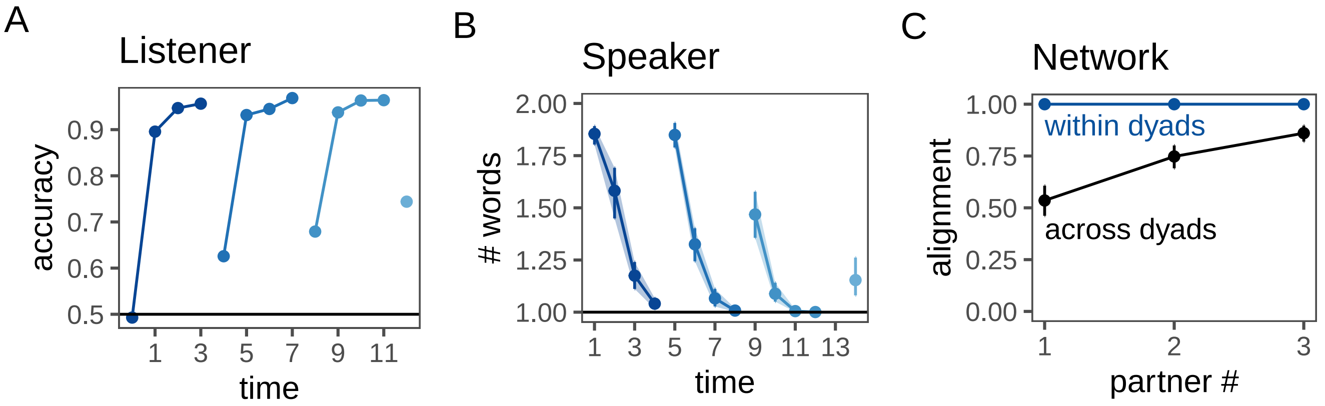

Figure 2: Model predictions across a series of different partners.

We define $P(D_k|\phi_k)$ as the probability of the data under a pragmatic listener L_1. We also use this RSA model to simulate the behavior of uncertain speakers S and listeners L. Utterances and object selections are sampled from the posterior predictive, marginalizing over lexical uncertainty.

Finally, we must specify the form of the lexical prior and a method to perform inference in this model. We assume Θ is a matrix with an entry for each utterance-object pair (u_i, o_j), and use independent Gaussian distributions for each $\Theta_{ij} \in \Theta$ as a hyper-prior. We then centered our partner-specific prior $\phi_{ij} \in \phi$ at the shared value for a particular partner:

$$\begin{array}{rcl} P(\Theta_{ij}) & \sim & \mathcal{N}(0,1) \\ P(\phi_{ij}|\Theta_{ij}) & \sim & \mathcal{N}(\Theta_{ij},1) \end{array}$$

The variances chosen in these priors represent assumptions about how far partner-specific priors can drift from the community-wide value.

For all simulations, we used the implementation of variational inference in WebPPL (Goodman & Stuhlmüller, n.d.). Variational methods transform probabilistic inference problems into optimization problems by approximating the true posterior with a parameterized family. Specifically, we make a *mean-field* approximation and assume that the full posterior can be factored into independent Gaussians for each random variable. We then optimize the parameters of these posterior Gaussians by minimizing the evidence lower bound (ELBO) objective (see Murphy, 2012 for more details). We run 50,000 steps of gradient descent on the first observation to obtain a posterior, compute the agent's marginal prediction for the next observation by taking the expectation over 50,000 samples from the posterior predictive, then continue running gradient descent on the same parameters after adding the new observation in the data.

Simulation 1: Listener accuracy across partners The key predictions of our model concern the pattern of generalization across partners. In our first simulation, we consider the partner-specificity of a *listener*'s expectations about which object is being referred to. To observe the model's behavior in the simplest case, we assume the speaker has a vocabulary of two single-word utterances $\{u_1, u_2\}$ with equal cost, and feed the listener the same utterance and feedback about the target object $(\{o_1, u_1\})$ on every trial. Instead of presenting these observations from a single partner, or randomly choosing a different partner on every trial, we swap in a new partner every block of 4 trials.

Our results are shown in Fig. 2A. The listener begins at chance due to its uninformative prior, but after observing several trials of evidence from the same partner, it rapidly infers the meaning of u_1 and learns to choose the true target with high accuracy. When a second partner is introduced, however, it reverts nearly to its original state. 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, this data is more parsimoniously explained with a partner-specific model. After observing multiple partners behave similarly, however, we find that this knowledge has gradually been incorporated into community-level expectations. This is evident in the much stronger initial expectations about meaning by its fourth partner ($\sim$ 75% accuracy vs. 50% with the first partner.)

Simulation 2: Speaker utterance length across partners Next, we examined our model's predictions about how a *speaker*'s referring expressions will change with successive listeners. While it has been frequently observed that messages reduce in length across repetitions with a single partner (Krauss & Weinheimer, 1964), and sharply revert back to longer utterances when a new partner is introduced (Wilkes-Gibbs & Clark, 1992), the key prediction distinguishing our model concerns behavior across subsequent partner boundaries. Complete-pooling accounts predict no change in number of words when a new partner is introduced. No-pooling accounts predict that roughly the same initial description length will re-occur with every subsequent interlocutor. Here we show that a partial pooling account predicts a more complex pattern of generalization.

To allow for reduction, we allowed a set of four primitive utterances, $\{u_1, u_2, u_3, u_4\}$, to be combined into two-word conjunctions, e.g. $\{u_1 + u_2, u_3 + u_4\}$. The meanings of

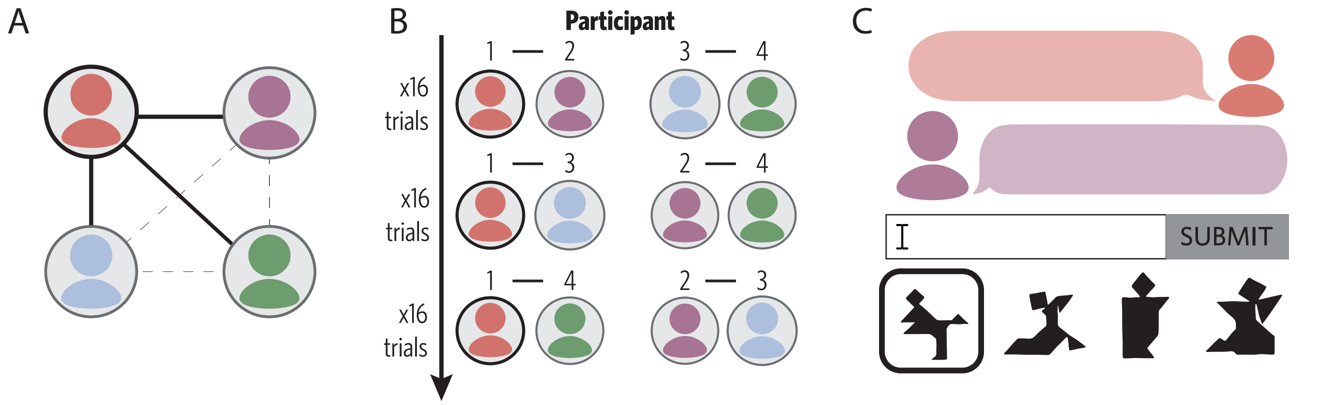

Figure 3: Experimental design. (A) Participants were placed in fully-connected networks of 4 and (B) played repeated reference games with each partner.

these two-word utterances were determined compositionally from the values of the primitive utterances[2]. Additionally, we placed a weakly biased initial prior over Θ: two of the utterances (u_1 and u_2) were assumed to apply more strongly to o_1 and the other two (u_3 and u_4) more strongly to o_2. This weak prior led the speaker to prefer conjunctions at the outset and thus allowed us to examine the speaker's shifting preference for conjunctions.

To focus on the speaker's behavior, we paired it with a fixed listener who always correctly selected the target, and ran 48 independent simulations. First, we find that successful interactions with the first partner become more efficient over trials as the model learns that the shorter utterances will be meaningful to their partner (e.g. the speaker begins to prefer either u_1 or u_2 to refer to o_1 instead of the conjunction; see Hawkins et al., 2017 for further interpretation of this effect). Second, we find that speakers revert back to a longer description at the first partner swap, just as our listener model did: evidence from a single partner is relatively uninformative about the community-level distribution. After interacting with several partners, however, speakers become increasingly confident that one or the other of the short labels is shared across the entire community, and are less likely to begin an interaction with a long utterance (Fig. 2B).

Simulation 3: Network convergence The first two simulations presented a single adaptive agent with a fixed partner to understand its gradient of generalization. In our final simulation, we test the consequences of the proposed hierarchical inference scheme for a network of interacting agents. From each individual agent's perspective, this simulation is identical to the earlier ones (i.e. a sequence of 3 different partners). Because all agents are simultaneously making inferences about the others, however, 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 used a round-robin scheme to schedule four agents into three blocks of interaction, with each agent taking turns in the speaker and listener roles, and again simulated 48 independent networks. We measured alignment by computing whether different pairs of agents produced the same one-word utterances as speakers. We compared the alignment between currently interacting agents (i.e. *within* a dyad) to the alignment between agents in the network who were not interacting (i.e. *across* dyads). During interaction with the first partner, alignment across dyads was roughly at chance, reflecting the arbitrariness of whether speakers reduce to u_1 or u_2. In the absence of hierarchical generalization, we would expect subsequent blocks to show similar chance levels, as partner-specific conventions would continually need to be re-negotiated from scratch. Instead, we find that alignment across dyads gradually increases, suggesting that partial pooling across partners leads to emergent consensus (Fig. 2C).

Behavioral experiment

To evaluate the qualitative predictions observed in our simulations, we designed a natural-language communication experiment following roughly the same network design. Instead of anonymizing partners, as in many previous empirical studies of convention formation, we divided the experiment into blocks of extended dyadic interactions with stable, identifiable partners (see Fay, Garrod, Roberts, & Swoboda, 2010; Garrod & Doherty, 1994 for similar designs). Each block was a full repeated reference game, where participants had to coordinate on an *ad hoc* convention, or *pact*, for how to refer to novel objects with their partner (Brennan & Clark, 1996). As the simulations demonstrated, our model predicts that these pacts will reset at partner boundaries, but that agents should be increasingly willing to transfer expectations from one part-

[2] We used the standard product T-norm semantics for conjunction in fuzzy logic, where values lie in the $[0,1]$ interval. Because we used a Gaussian prior with support over the real numbers, we first used a logistic function to map primitive values to the unit interval, and a logit function to map the product back to the original domain.

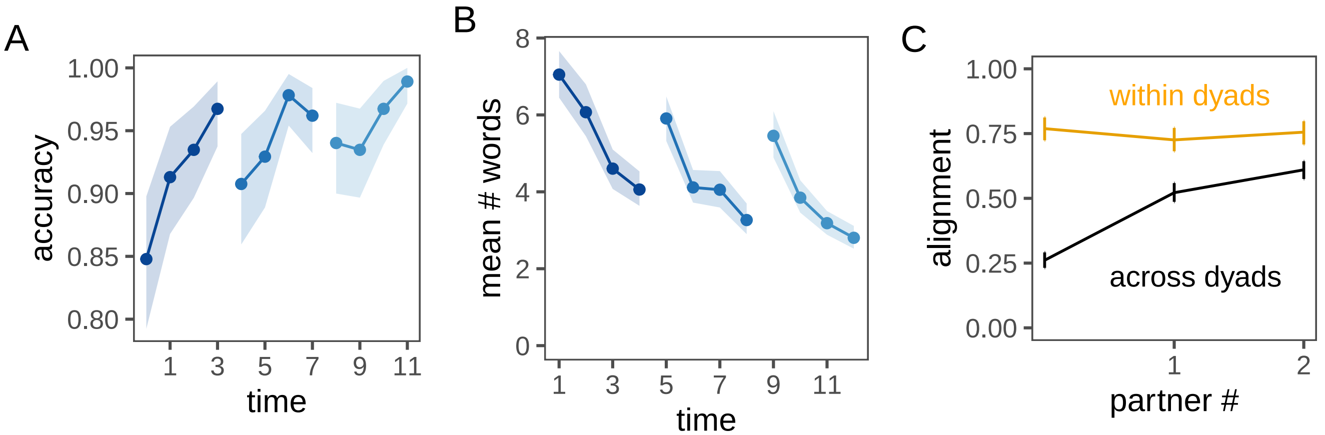

Figure 4: (A)Increase in accuracy across partners, (B) reduction in number of words across partners, (C) network convergence.

ner to another in their community.

Participants We recruited 92 participants from Amazon Mechanical Turk to play a series of interactive, natural-language reference games.

Stimuli and procedure Each participant was randomly assigned to one of 23 fully-connected networks with three other participants as their 'neighbors' (Fig. 3A). Each network was then randomly assigned one of three distinct "contexts" containing abstract tangram stimuli taken from Clark and 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. 3B). The trial sequence for each reference game was composed of four repetition blocks, where each target appeared once per block. After completing sixteen trials with one partner, participants were introduced to their next partner and asked to play the game again. This process repeated until each participant had partnered with all three neighbors. Because some pairs within the network took longer than others, we sent participants to a temporary waiting room if their next partner was not ready.

Each trial proceeded as follows. First, one of the four tangrams in the context was highlighted as the *target object* for the "speaker." They were instructed to use a chatbox to communicate the identity of this object to their partner, the "listener" (see Fig. 3C). The listener could reply freely through the chatbox but was asked to 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 representing their current partner to emphasize that they were speaking to the same partner for an extended period.

Results

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

Listener accuracy We first examined changes in listener accuracy over different partners, where accuracy is defined as the proportion of trials where the target was correctly selected (see Fig. 4A). In particular, our partial pooling model predicts (1) gains in accuracy within each partner and (2) drops in accuracy at partner boundaries, but (3) overall improvement in initial interactions with successive partners. To test the first prediction, we constructed a logistic mixed-effects regression predicting trial-level listener responses. 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 accuracy improved over successive repetitions with every partner, $b = 0.69$, $t = 3.87$, $p < 0.001$.

To test changes at partner boundaries, we constructed another regression model. We coded the blocks before and after each partner swap and compared the final repetition block with one partner against the first repetition block with the next. Because partner roles were randomized for each game, the same participant often did not 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). We found that across the two partner swaps, accuracy dropped significantly, $b =$ -1.56, $t =$ -2, $p =$ 0.045, reflecting the partner-specificity of meaning. Finally, to test whether performance in the initial repetition block improves with subsequent partners, we examined the simple effect of partner number on first repetition block. As predicted, we found a significant improvement in performance, $b = 0.57$, $t = 2.72$, $p = 0.007$, suggesting that listeners are bringing increasingly well-calibrated expectations into interactions with novel neighbors.

Speaker utterance length Next, as a measure of coding efficiency, we calculated the raw number of words produced by a speaker on each trial. We then tested analogs of the same three predictions we tested in the previous section: speakers should *reduce* utterance length with each partner and revert to longer utterances at partner boundaries, but become gradually more willing to use shorter referring expressions with successive partners. We tested these predictions using the same mixed-effects models, but using (log) utterance length as a continuous DV (see Fig. 4B). First, we found that speaker used fewer words over the course of interaction with every partner, $b = -0.19$, $t = -9.88$, $p < 0.001$. Second, we found that length increased across partner-boundaries, $b = 0.43$, $t = 4.4$, $p < 0.001$, indicating speaker sensitivity to different partners. Finally, we found an incremental decrease in the lengths of *initial descriptions* as speakers interacted with more partners on their network,$b = -0.2$, $t = -6.07$, $p < 0.001$.

Network convergence While coarse signatures of accuracy and utterance length are consistent with the predictions of our partial pooling model, it is possible that the network as a whole still fails to coordinate. In this section, we examine the actual *content* of pacts and measure changes in alignment across the network. Specifically, we extend the 'exact matching' measure of alignment used in Simulation 3 to messier natural language production by examining the *intersection* of words produced by different speakers, excluding common stop words (e.g. 'the', 'both') to emphasize the core conceptual content of the pact. We then constructed a mixed-effects logistic regression predicting whether this intersection was non-empty for each pair of utterances produced by speakers.

The main comparison of interest was between currently interacting participants (within dyad), and participants who are not interacting (across dyad). Because the latter group is not aware of one another, if they nonetheless happen to tacitly be aligned, it provides evidence of network-wide coordination. Our prediction thus concerns the interaction between pair type and partner number: within-pair alignment should stay consistently high while (tacit) alignment between non-interacting pairs will gradually increase as participants interact with more partners. We included these variables and their interaction as fixed effects, along with random intercepts for each tangram and maximal random effects for each network (i.e. intercept, both main effects, and the interaction). We found a significant interaction ($b = -0.85$, $t = -5.69$, $p < 0.001$; see Fig. 4C), indicating that although different pairs in a network may initially use different labels, some of these labels begin to spread through the network over subsequent interactions.

Discussion

How do community-level conventions emerge from local interactions? In this paper, we suggested a partial pooling account of convention formation, formalized as a hierachical Bayesian model, where conventions represent the shared structure that agents can "abstract away" from partner-specific interactions. Unlike complete pooling accounts, this model allows for partner-specific common ground to override existing community-wide expectations given sufficient experience with a partner, or in the absence of strong conventions. Unlike no-pooling accounts, it allows networks to gradually converge on more efficient and accurate initial expectations about new partners. We conducted a series of simulations demonstrating key model predictions about generalization behavior, and evaluated these predictions in a natural-language communication experiment where participants interacted with different neighbors on their network.

Hierarchical Bayesian models have several other properties of theoretical interest for convention formation that may be useful to evaluate in future work. First, they offer a "blessing of abstraction" (Goodman, Ullman, & Tenenbaum, 2011), where community-level conventions may be learned even with relatively sparse input from each partner, as long as there is not substantial variance in the population. Second, they are more robust to deviations than complete-pooling models relying on a fixed set of memory slots or a single mental lexicon. This robustness is due to their ability to 'explain away' outliers with partner-specific models without their community-level expectations being affected. They can therefore explain how speakers are able to emerge with conventional knowledge intact from an extended series of communicative interactions where they have adapted to children or non-native speakers. Finally, the deep connection between hierarchical Bayesian models and accounts of *meta-learning*, or learning to learn (e.g. Grant, Finn, Levine, Darrell, & Griffiths, 2018), provides a useful set of tools to analyze conventions as the result of agents solving a meta-learning problem, adapting to each partner along the way.

The current work captures and quantifies incremental convergence within communities of four unfamiliar English speakers towards a set of shared language conventions. We recognize that real-world communities are more complex than this, however, as each speaker takes part in a number of subcommunities which vary in size and overlap. For example, we use partially distinct conventions depending on whether we are communicating with psychologists, friends from high school, bilinguals, or children, and we are able to comprehend certain conventions that we do not use ourselves. For future work using hierachical Bayesian models to address the full scale of an individual's network of communities, additional social knowledge about these communities must be learned and represented in the generative model (e.g. Gershman, Pouncy, & Gweon, 2017). Our results are a promising first step, providing evidence that hierarchical generalization may be a foundational cognitive building block for establishing conventionality at the group level.

References

Baronchelli, A. (2018). The Emergence of Consensus. *Royal Society Open Science*, *5*(2).

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

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

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

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.

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

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

Davidson, D. (1986). A nice derangement of epitaphs. *Philosophical Grounds of Rationality: Intentions, Categories, Ends*, *4*, 157–174.

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.

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

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

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

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

Gershman, S. J., Pouncy, H. T., & Gweon, H. (2017). Learning the Structure of Social Influence. *Cognitive Science*, *41*. `http://doi.org/10.1111/cogs.12480`

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. (n.d.). The design and implementation of probabilistic programming languages. Retrieved from `http://dippl.org`

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

Grant, E., Finn, C., Levine, S., Darrell, T., & Griffiths, T. (2018). Recasting gradient-based meta-learning as hierarchical bayes. *arXiv Preprint arXiv:1801.08930*.

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

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.

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.

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

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).

Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. MIT press.

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

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.

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

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