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

Robert D. Hawkins [1], Noah D. Goodman[2], Adele E. Goldberg[1], Thomas L. Griffiths[1,3]
[1]Department of Psychology, Princeton University ({rdhawkins,adele,tomg}@princeton.edu)
[2]Department of Psychology and Computer Science, Stanford University (ndgoodman@stanford.edu)
[3]Department of Computer Science, Princeton University

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 generalize meanings that seem to be shared across partners. To evaluate 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 and find key signatures of both partner-specificity and community convergence that distinguish our model from alternatives. These results suggest that local partner-specific adaptation is not only compatible with the formation of community-wide conventions, but may facilitate it when coupled with a powerful inductive mechanism.

Keywords:

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 (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 (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 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 (Davidson, 1986; 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 transferred 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 pairs, but also gradual generalization of these local conventions 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 hierarchical inference mechanism.

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 linguistic meaning can be represented probabilistically: agents have uncertainty about the form-meaning mappings their current partner is using (Bergen, Levy, & Goodman, 2016). In our hierarchical model, this 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 meaning used by partner k (see Fig. 1). Given observations D_k from interactions with partner k, the agent updates their beliefs about the latent system of meaning using Bayes rule:

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

This joint inference decomposes the 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 captures the idea that different partners may share aspects of meaning in common. In the absence of strong evidence that partner-specific language use departs from this common structure, the agent ought to regularize toward background knowledge of the population's conventions. The likelihood represents predictions about how a particular partner will use language under different systems of meaning.

The joint posterior over meanings in Eq. 1 has two consequences for convention formation. First, it allows agents to maintain idiosyncratic partner-specific expectations ϕ_k by marginalizing over community-level uncertainty:

$$P(\phi_k|D_k) = \int_\Theta P(D_k|\phi_k)P(\phi_k|\Theta)P(\Theta)d\Theta \quad (2)$$

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

$$P(\Theta|D) = P(\Theta)\int_\phi P(D_k|\phi_k)P(\phi_k|\Theta)d\phi \quad (3)$$

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 hierarchical inference is sometimes referred to as "sharing of strength" or "partial pooling" (Gelman & Hill, 2006) because it smoothly integrates aggregated data (as in complete-pooling models) with domain-specific knowledge (as in no-pooling models).

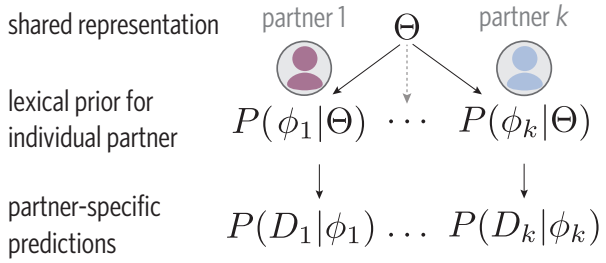

Figure 1: Schematic of hierachical Bayesian model.

Model simulations

We investigate the qualitative predictions of our partial-pooling model, and compare it to complete-pooling and no-pooling models, under three increasingly complex scenarios. In all of these scenarios, speaker and listener agents play a reference game with a set of two objects $O = \{o_1, o_2\}$. On each trial, one of these objects is privately identified to the speaker as the *target*. They must then select from a set of utterances $\mathcal{U} = \{u_0, \ldots, u_j\}$ to convey the identity of the target to the listener. Upon hearing this utterance, the listener must select which of the objects they believe to be the target. Both agents then receive feedback about the selection and 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.

In this reference game setting, we can explicitly specify the likelihood and prior terms in Eq. (1). We consider a likelihood given by the Rational Speech Act (RSA) framework, which formalizes the Gricean assumption of cooperativity (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 the intended target directly using a softmax over the parameterized lexical meaning function $\mathcal{L}_{\phi_k}$. We assume, for simplicity, that ϕ_k is a $|O| \times |\mathcal{U}|$ real-valued matrix with entries for each utterance-object pair, and $\mathcal{L}_{\phi_k}(o,u)$ simply looks up the entry for (o,u). This model can be formally specified as follows:

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

$c(u)$ is the cost of producing u and w_I and w_C are free parameters controlling the relative weights on the informativity and parsimony, respectively[1]. Note that under each value of ϕ_k,

[1]Throughout our simulations, we set $w_I = 11$, $w_C = 7$. A grid search over parameter space revealed other regimes of behavior, but we leave broader exploration of this space to future work.

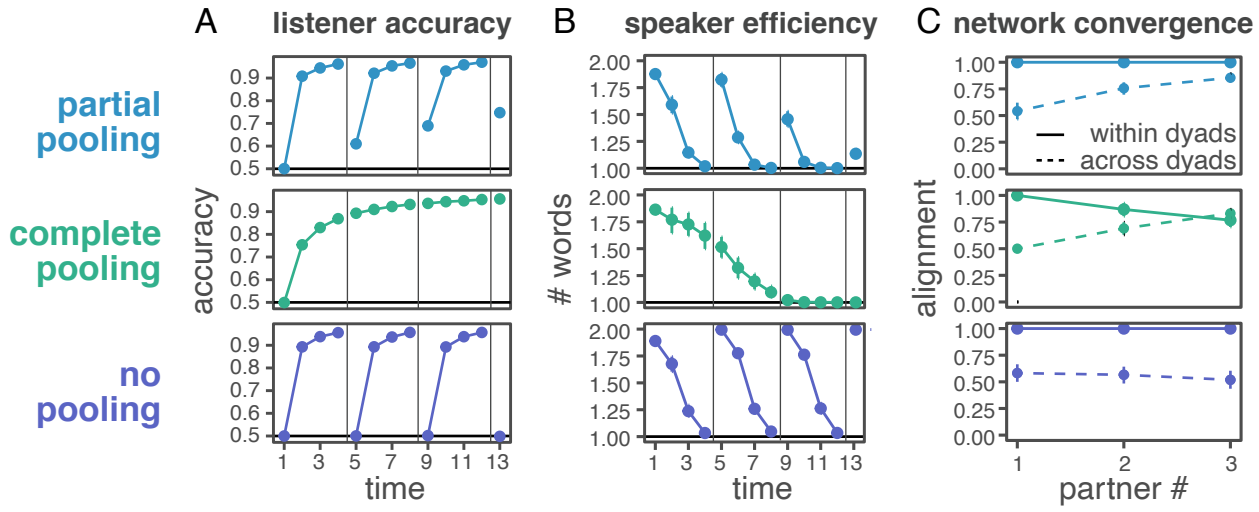

Figure 2: Simulation results for (A) listener accuracy, (B) speaker efficiency, and (C) network convergence across three partners.

the S_1 and L_1 functions assign a probability to each word or object that one's partner has chosen, thus yielding the likelihood of the full set of observations $P(D_k|\phi_k)$. In addition to using the RSA likelihood for updating an agent's beliefs about Θ and ϕ_k, we can use the same functions to simulate the choices of speaker and listener agents on a particular trial. That is, we sample from the posterior predictive, marginalizing over the agent's current beliefs about ϕ_k:

$$\begin{aligned} L(o|u) &\propto \int_{\phi_k} P(\phi_k|D_k) S_1(u|o,\phi_k) \\ S(u|o) &\propto \exp\{\int_{\phi_k} w_I \cdot P(\phi_k|D_k) \log L_1(o|u,\phi_k) - w_C \cdot \mathrm{c}(u)\} \end{aligned}$$

Finally, we must specify the form of the hierarchical lexical prior and a method to perform inference in this model. The hyper-prior for Θ is given by independent Gaussian distributions for each matrix entry $\Theta_{ij} \in \Theta$. We then center the partner-specific prior $\phi_{ij} \in \phi$ at the corresponding value Θ_{ij}:

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

These priors represent assumptions about how far partner-specific learning can drift from the community-wide value.

For all simulations, we used variational inference as implemented 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 (Murphy, 2012; Ranganath, Gerrish, & Blei, 2013). We run 50,000 gradient steps 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 resume gradient descent on the resulting outcome.[2]

Simulation 1: Listener accuracy across partners The key predictions of our partial-pooling 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 possible case, we assume the speaker has a vocabulary of two single-word utterances $\{u_1, u_2\}$ with equal cost and produces the same utterance for the same target object ($\{o_1, u_1\}$) on every trial. We introduce a new partner every 4 trials.

The probability the listener assigns to the target on each trial is shown for the different models in Fig.˜2A. Under the partial-pooling model (top row), the listener agent 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 higher accuracy. When a second partner is introduced, however, the agent's expectations revert nearly to their original state, unlike the complete-pooling model (middle row). 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, its observations are more parsimoniously explained at the partner-specific level. After observing multiple partners use u_1 similarly, however, we find that this knowledge has gradually been incorporated into community-level expectations. This is evident in much stronger initial expectations when introduced to the fourth partner ($\sim$ 75% accuracy vs. 50% with the first partner), unlike the no-pooling model (bottom row).

[2]These details were chosen to ensure high-quality estimates of the model's behavior; see Discussion for other algorithmic possibilities.

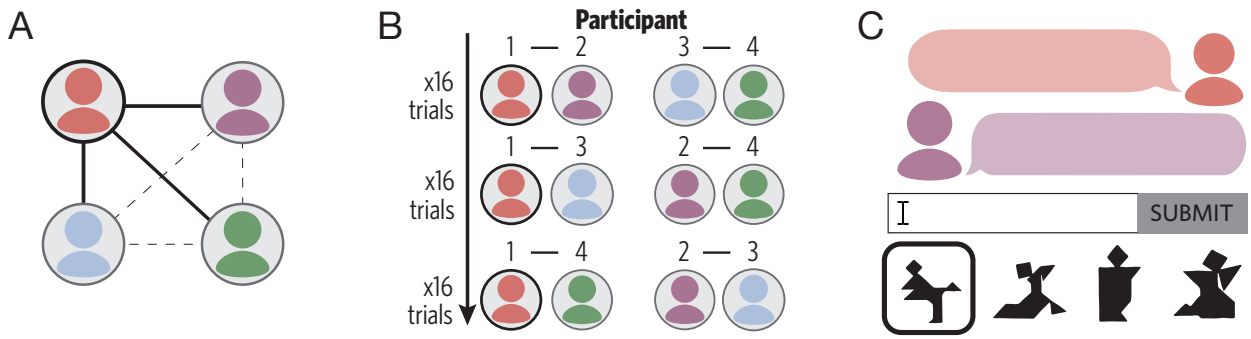

Figure 3: In our experiment, (A) participants were placed in fully-connected networks of 4, (B) paired in a round-robin schedule with each neighbor, and (C) played a series of repeated reference games using tangram stimuli.

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 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.˜2B, middle row). No-pooling accounts predict that roughly the same initial description length will re-occur with every subsequent interlocutor (Fig.˜2B, bottom row). Here we show that a partial pooling account predicts a more complex pattern of generalization.

We allowed a set of four primitive utterances, $\{u_1, u_2, u_3, u_4\}$, to be combined into conjunctions, e.g. $\{u_1 + u_2, u_3 + u_4\}$, which are assumed to have a higher utterance cost. The meanings of these conjunctions were determined compositionally from the values of the primitive utterances[3]. Speakers do not typically begin at chance over their *entire* vocabulary, so we introduced a weakly biased prior for Θ: two of the primitive utterances initially applied more strongly to o_1 and the other two more strongly to o_2. This weak bias leads to a preference for conjunctions at the outset and thus allows us examine reduction.

We paired the speaker model with a fixed listener who always selected the target, and ran 48 independent simulations. First, we found that descriptions become more efficient over interaction with a single partner: the model becomes more confident that shorter utterances will be meaningful, so the marginal informativity provided by the conjunction is not worth the additional cost (see Hawkins et al., 2017). Critically, we find that the speaker model reverts back to a longer description at the first partner swap: evidence from one partner is relatively uninformative about the community. After interacting with several partners, however, it becomes more likely that one of the short labels is shared across the entire community, and the model is correspondingly more likely to begin a new interaction with it (Fig.˜2B, top row).

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 simulated 48 networks. We measured alignment by computing whether different agents produced the same one-word utterances. We compared the alignment between currently interacting agents (i.e. *within* a dyad) to those who were not interacting (i.e. *across* dyads). Alignment across dyads was initially near chance, reflecting the arbitrariness of whether speakers reduce to u_1 or u_2. In the absence of hierarchical generalization (Fig.˜2C, bottom row), subsequent blocks remain at chance, as conventions need to be re-negotiated from scratch. Under complete pooling (Fig.˜2C, middle row), agents persist with mis-calibrated expectations learned from previous partners rather than quickly adapting to their new partner, and *within-dyad* alignment deteriorates. By contrast, alignment across dyads gradually increases for the partial pooling model, suggesting that hierarchical inference leads to emergent consensus (Fig. 2C).

Behavioral experiment

To evaluate the predictions observed in our simulations, we designed a natural-language communication experiment following roughly the same network design. Instead of

[3]We used a standard product T-norm for conjunction. Because our values come from a Gaussian prior and the T-norm is defined over $[0, 1]$, we used logistic and logit function to map values to the unit interval and back.

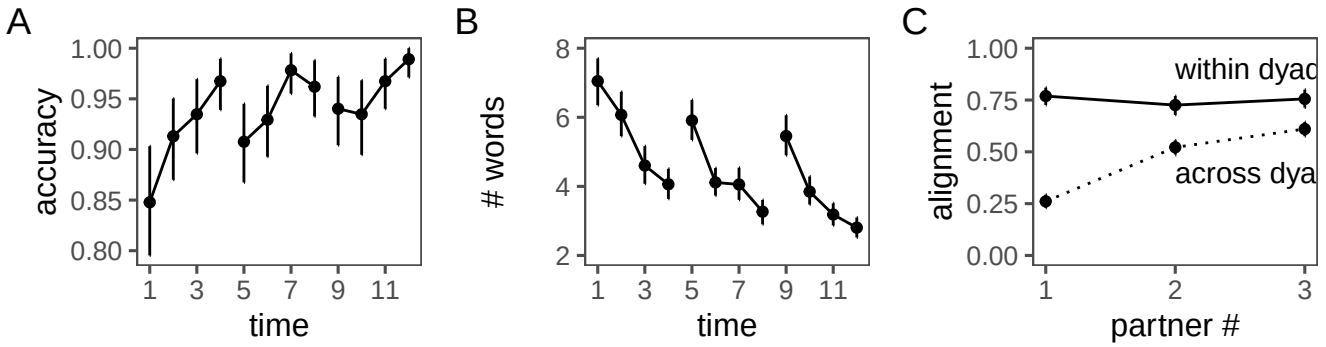

Figure 4: Results from behavioral experiment. (A) listener accuracy improved across partners, (B) reduction in number of words across partners, (C) network convergence.

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 *ad hoc* conventions for how to refer to novel objects with their partner (Brennan & Clark, 1996). Our partial-pooling model predicts that these conventions will partially reset at partner boundaries, but that 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.

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 the proportion of correct listener selections. 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, $b = 0.69$, $z = 3.87$, $p < 0.001$.

To test changes at partner boundaries, we constructed another regression model. We coded the repetition blocks immediately before and after each partner swap, and included this as a categorical fixed effect. 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, $z =$ -2, $p < 0.05$, reflecting partner-specificity of meaning. Finally, to test whether performance improves for the *very first* interaction with each new partner, before observing any partner-specific information, we examined the simple effect of partner number on the trials immediately after the partner swap ($t = \{1, 5, 9\}$). As predicted, we found a significant improvement in performance, $b = 0.57$,

$z = 2.72$, $p < 0.01$, suggesting that listeners are bringing increasingly well-calibrated expectations into interactions with novel neighbors (see Fig.˜4A).

Speaker utterance length Next, as a measure of coding efficiency, we calculated the raw length (in words) of the utterance produced on each trial. We then tested analogs of the same three predictions we tested in the previous section using the same mixed-effects models, but using a linear regression on the continuous measure of efficiency instead of accuracy (see Fig.˜4B).[4] We found that speakers reduced utterance length with every partner, $b =$ -0.19, $t(34) =$ -9.88, $p < 0.001$, increased length across partner-boundaries, $b =$ 0.43, $t(22) = 4.4$, $p < 0.001$, and decreased the length of their *initial descriptions* as they interacted with more partners on their network,$b = -0.2$, $t(516.5) = -6.07$, $p < 0.001$ (see Fig.˜4B).

Network convergence In this section, we examine the actual *content* of pacts and test whether these coarse signatures of generalization actually lead to increased alignment across the network, as predicted. Specifically, we extend the 'exact matching' measure of alignment used in Simulation 3 to natural language production by examining whether the *intersection* of words produced by different speakers was non-empty[5]. As in our simulation, the main comparison of interest was between currently interacting participants and participants who are not interacting: we predicted that within-pair alignment should stay consistently high while (tacit) alignment between non-interacting pairs will increase. We thus constructed a mixed-effects logistic regression including fixed effects of pair type (within vs. across), partner number, and their interaction. We included random intercepts at the tangram level and maximal random effects at the network level (i.e. intercept, both main effects, and the interaction). As predicted, we found a significant interaction ($b =$ -0.85, $z =$ -5.69, $p < 0.001$; see Fig. 4C). Although different pairs in a network may initially use different labels, these labels begin to align over subsequent interactions.

Discussion

How do community-level conventions emerge from local interactions? In this paper, we proposed a partial-pooling account, formalized as a hierarchical Bayesian model, where conventions represent the shared structure that agents "abstract away" from partner-specific interactions. Unlike complete-pooling accounts, this model allows for partner-specific common ground to override community-wide expectations given sufficient experience with a partner, or in the absence of strong conventions. Unlike no-pooling accounts, it allows networks to converge on more efficient and accurate expectations about novel partners. We conducted a series of simulations demonstrating the model's generalization behavior, and evaluated these predictions with a natural-language communication experiment on a small network.

Hierarchical Bayesian models have several other properties of theoretical interest for convention formation. First, they offer a "blessing of abstraction" (Goodman et al., 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 partner-specific deviations from conventions (e.g. interactions with children or non-native speakers) than complete-pooling models relying on a fixed set of memory slots or a single mental 'inventory.' This robustness is due to their ability to 'explain away' outliers without community-level expectations being affected. Finally, the deep connection between hierarchical Bayesian models and accounts of *meta-learning*, or learning to learn (Grant et al., 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.

Real-world communities are much more complex than the simple networks we considered: each speaker takes part in a number of overlapping subcommunities. For example, we use partially distinct conventions depending on whether we are communicating with psychologists, friends from high school, bilinguals, or children. For future work using hierarchical Bayesian models to address the full scale of an individual's network of communities, additional social knowledge about these communities must be learned and represented in the generative model (e.g. Gershman et al, 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.

Acknowledgments

This work was supported by NSF SPRF-FR grant #1911835 to RDH, TLG, and AEG, and CV Starr Fellowship to RDH.

All code and materials for simulations, data analyses, and web experiment available at:
`https://github.com/hawkrobe/conventions_model`

References

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).

Brennan, S., & Clark, H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(6), 1482–1493.

[4] We log-transformed utterance lengths for stability.

[5] We excluded a list of common stop words (e.g. 'the', 'both') to focus on the core conceptual content of pacts; using the size of the intersection instead of the binary variable yielded similar results.

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, England: Cambridge University Press.

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), 351–386.

Garrod, S., & Doherty, G. (1994). Conversation, coordination & convention: An empirical investigation of how groups establish linguistic conventions. *Cognition*, *53*(3), 181–215.

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

Gershman, S., Pouncy, H., & Gweon, H. (2017). Learning the Structure of Social Influence. *Cognitive Science*, *41*.

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), 110–119.

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. 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*.

Kleinschmidt, D., & Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review*, *122*(2), 148–203.

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.

Ranganath, R., Gerrish, S., & Blei, D. M. (2013). Black box variational inference. *arXiv Preprint arXiv:1401.0118*.

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.