Respect the code: Speakers expect novel conventions to generalize within but not across social group boundaries

Robert D. Hawkins, Irina Liu, Adele E. Goldberg, Thomas L. Griffiths
Department of Psychology, Princeton University
{robertdh,irinal,adele,tomg}@princeton.edu

Abstract

Speakers use different language to communicate with partners in different communities. But how do we learn and represent which conventions to use with which partners? In this paper, we argue that solving this challenging computational problem requires speakers to supplement their lexical representations with knowledge of social group structure. We formalize this idea by extending a recent hierarchical Bayesian model of convention formation with an intermediate layer explicitly representing the latent communities each partner belongs to, and derive predictions about how conventions formed within a group ought to extend to new in-group and out-group members. We then present evidence from two behavioral experiments testing these predictions using a minimal group paradigm. Taken together, our findings provide a first step toward a formal framework for understanding the interplay between language use and social group knowledge.

keywords: conventions, communication, social cognition

There is tremendous variation in the linguistic conventions used for communication by different communities (Gumperz, 1982; Eckert, 2012). This variation manifests most strikingly in the hundreds of mutually unintelligible language families currently in use around the world (Katzner & Miller, 2002). Yet even among different speakers of a single language (e.g. English), there exist communities that remain nearly unintelligible to one another. The same acronym may have entirely different meanings in different scientific journals, and the slang used by one generation may be foreign to the previous one (Eble, 1996; Partridge, 2006). Such variation creates a challenging computational problem for speakers: in an increasingly interconnected world, individuals are likely to belong not just to one language community but many, spanning different professional, ethnic, and interest-based groups. What cognitive abilities allow speakers to successfully navigate this landscape of inter-group variability?

An extensive body of work in sociolinguistics has focused on one solution, known as *code-switching* (DeBose, 1992; Auer, 2013; Gardner-Chloros, 2009), the ability to retrieve and use different conventions with different partners in different contexts. For instance, when a scientist presents their work to other scientists, they may use efficient technical shorthand that they would avoid when talking to their non-expert friends. Recent accounts of language use have suggested that such flexibility may be supported by the ability to encode partners as latent contexts representing which conventions are expected to be shared (Brown-Schmidt, Yoon, & Ryskin, 2015), and the ability to generalize these latent contexts appropriately to new individuals (Hawkins, Goodman, Goldberg, & Griffiths, 2020). In practice, however, these accounts have tended to focus on common ground within the most minimal possible communities (the agent and exactly one partner) or the most maximal (the entire population).

A core problem facing these accounts, then, is that conventions are often inextricable from knowledge about the latent structure of the social world: between partners and the population lay many intermediate communities that must be learned and represented (Gershman, Pouncy, & Gweon, 2017; Lau, Pouncy, Gershman, & Cikara, 2018). In other words, the prior expectations that guide communication with a new partner should neither be a blank slate nor a copy of global expectations but should instead be based on inferences about latent group membership. For example, in a compelling empirical demonstration of these group-based inferences, Isaacs and Clark (1987) paired participants who had previously lived in New York City with those who had never been there, and asked them to take turns referring to images of landmarks in the city (e.g. the "Rockefeller Center"). After a handful of utterances from a novel partner, participants could infer whether they were playing with an expert (i.e. in-group member of the New York community) or a novice (i.e. an out-group member) and modified their descriptions accordingly.

In this paper, we propose a computational model that aims to both explain the ability to *acquire* group-specific conventions and to deploy them appropriately in conversation with different partners. We evaluate this model's predictions using empirical data from a minimal group paradigm (Kerr & Smith, 2016; Tajfel, 1982) implemented with a networked communication task (Experiment 1). This task aimed to examine one of the weakest conditions under which conventions may be expected to depend on a partner's social group. We arbitrarily assigned participants to either a 'blue' community or a 'red' community and had them take turns describing ambiguous tangram objects in interactions with different partners in their own community. At the end of the experiment, we asked each participant to produce descriptions of the same objects for members of their in-group or for members of the out-group. Finally, we showed these descriptions to naive participants to evaluate the transparency of descriptions produced for in-group vs. out-group members (Experiment 2). In both cases, we found small effects of intended audience, suggesting that speakers are sensitive to social group structure when forming and using conventions.

Reasoning about social group structure

Hawkins et al. (2020) recently proposed an account of convention formation based on the idea that communicative agents maintain uncertainty about how different partners will use language, and update their expectations based on evidence from communicative interactions. Rather than maintaining entirely disconnected expectations for each partner (a *no-pooling* model) or using exactly the same expectations for every partner (a *complete-pooling* model), this account was formalized as a hierarchical Bayesian model (see Tenenbaum, Kemp, Griffiths, & Goodman, 2011), where agents maintain partner-specific expectations while abstracting away what is shared in common partners.

Our primary theoretical aim is to integrate this hierarchical account of communicative expectations into a unified framework with recent models of social structure learning. Following Gershman et al. (2017), we assume each partner i has some (latent) group membership z_i, where members of a group are assumed share some attributes θ_{z_i} in common[1]. We then assume these groups are represented at an intermediate layer in a generative model of a partner's behavior (see Fig. 1). The parameters of each possible latent social group θ_{z_i} are sampled from an overall population distribution, $P(\theta_{z_i}|\Theta)$, where $P(\Theta)$ represents the highest-level uncertainty about the population-level parameters. Meanwhile, the lexicons used by individuals within each group are drawn from their respective group-specific distributions, $P(\phi_i|\theta_{z_i})$. This three-layer *community-sensitive* structure contrasts with a *community-free* model, where individuals are drawn directly from the population-level distribution.Otherwise, the two models proceed similarly to allow agents to dynamically update their beliefs by inverting this model conditioned on data. Concretely, as an agent interacts with each partner i, they make observations D_i about their partner's language use in context. These observations may be used to update their joint beliefs over parameters at each layer, using Bayes rule,

$$\begin{aligned} P(\Theta,\theta_{z_i},\phi_i|D_i) &\propto P(D_i|\phi_i,\theta_{z_i},\Theta)P(\phi_i,\theta_{z_i},\Theta) \\ &= P(D|\phi_i)P(\phi_i|\theta_{z_i})P(\theta_{z_i}|\Theta)P(\Theta). \end{aligned}$$

The key observation is that variability at each layer of the generative model modulates the strength of the inferences that can be made as parameters recede from the observed data. For example, after coordinating with a single partner, an agent is able to form strong expectations about how that specific partner will use language in the future, allowing more effective communication. However, the same data does *not* license strong inferences about whether that partner is representative of their group, or whether their group is representative of the general population. Most importantly for the mechanisms underlying code-switching, this model predicts that when conventions form within a group, they will initially be limited to in-group members.

[1]In principle, memberships z_i are unknown and must be inferred alongside the properties of each group. We restrict our analysis to the case where memberships are already known.

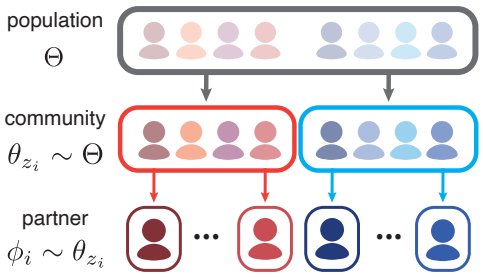

Figure 1: We consider a hierarchical model of convention formation that represents not only partner-specific common ground but also the latent structure of social communities.

We make these theoretical observations explicit by considering a group of agents communicating with one another in a fully-connected network. We begin by reproducing the simulations reported by Hawkins et al. (2020) to explicitly compare the predictions of a *community-free* model against our new *community-sensitive* model. In these simulations, we initialize four agents, assign them all to the same group (e.g. the 'red' team), and pair them up in a series of round-robin *repeated reference games* (see Fig. 2). One agent is assigned to the *speaker* role and shown a context of objects $\mathcal{C}$ with one object $o^* \in \mathcal{C}$ indicated as the target. Their objective is to choose an utterance $u \in \mathcal{U}$ that allows their partner, the *listener*, to accurately choose the target. In our minimal setting, we set $|\mathcal{C}| = 2$ and give the speaker a vocabulary of four words, which can be concatenated to form longer utterances.

We specify agents' referential language behavior using the Rational Speech Acts (RSA) framework (Goodman & Frank, 2016) for consistency with previous work. The speaker is assumed to choose utterances by balancing the expected communicative success against the utterance's cost, where longer utterances are assumed to be more costly. Critically, an utterance's expected communicative success depends on the lexicon ϕ the listener is using to interpret the utterance. We define the listener to select between objects using a softmax distribution: $P_L(o|u,\phi) \propto e^{\phi[u,o]}$, where ϕ is a real-valued 2×4 matrix. The lexicon ϕ is precisely the set of parameters that the speaker maintains uncertainty about, via the distribution $P(\phi_i|D)$; we therefore assume the speaker marginalizes over their current beliefs to choose an utterance:

$$P_S(u|o,\phi) \propto \exp\{\int_{\phi_i} P(\phi|D) w_I \cdot \ln P_L(o|u,\phi) - w_c \cdot c(u)\}$$

where w_I and w_c control how strongly informativity and utterance cost $c(u)$ are weighted in the utility. We set $w_I = 12$, $w_c = 7$, and use independent Gaussian distributions as priors for each cell of the lexicon matrix $\phi_{ij}^{(k)}$. These distributions are centered at the corresponding value of the group-level matrix, which in turn are centered at the value of the population-level matrix (for further details, see Hawkins et al., 2020):

$$P(\Theta_{ij}) = \mathcal{N}(0,1),\ P(\theta_{ij}^{(z_k)}) = \mathcal{N}(\Theta_{ij},1),\ P(\phi_{ij}^{(k)}) = \mathcal{N}(\theta_{ij}^{(z_k)},1)$$

Results are shown for 33 networks in Fig. 3A.

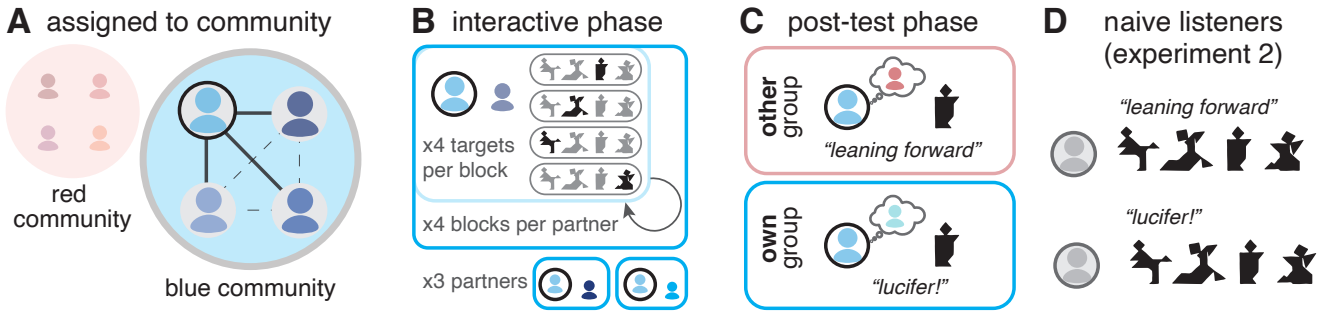

Figure 2: *Design and procedure for simulations and experiments.* (A) In our simulations and in Experiment 1, participants were assigned to a red or blue community and (B) played a series of reference games with their neighbors. (C) At the end of the task, they were asked to produce a description for a novel partner who belonged to their same group, or who belonged to the other group. (D) In Experiment 2, a group of naive participants were asked to select the intended target based on the description.

Experiment 1: Generalizing to in-group vs. out-group members

To test these predictions, we evaluated inter- and intra-group generalization by introducing a minimal group paradigm for a referential communication task. Our key prediction concerned responses in a post-test phase, where we asked participants to produce descriptions of each object for a new member of their own community as well as a new member of the other community. Therefore, the descriptions they provide in the post-test ought to be shorter in description length for a novel member of their own group than for a novel member of the other group.

Methods

Participants We recruited 272 participants from Prolific and connected them in groups of four using a reactive web app built with Empirica (Almaatouq et al., 2021). All participants were pre-screened as fluent (but not necessarily 'native') English speakers. We deliberately recruited more participants than required for each network, to ensure each network had enough participants to begin; over-subscribed participants were paid the base rate of $4.00. Active participants could receive up to $2.24 in additional performance bonuses. After excluding incomplete games, where at least one participant disconnected prior to competition of the full task, we were left with complete data from 33 groups, consisting of 132 unique participants.

Design & Procedure Each group of four participants was randomly assigned one of two possible team colors ('red' or 'blue') and one of two possible object sets containing four tangram stimuli from Clark and Wilkes-Gibbs (1986, see Fig. 2A). The experiment was structured into a series of dyadic repeated reference games using these stimuli as targets. Partner pairings were determined by a round-robin schedule, such that every participant had an extended interaction with each of their neighbors in a private room (Fig. 2B). The trial sequence was composed of four repetition blocks per partner, where each target appeared exactly once per block. Participants swapped speaker and listener roles at the beginning of each block. After completing sixteen trials with one partner, participants were introduced to their next partner. To emphasize the continuity of interaction with the same partner in a room, as well as each partner's group membership, we graphically represented participants as avatars using their team color (i.e. shades of blue or red).

Each trial proceeded as follows. First, one of the four tangrams in the context was highlighted as the *target object* for the current speaker in the room. They were instructed to use a chatbox to communicate the identity of this object to their partner, the listener. The two participants were able to communicate freely through the chatbox until the listener decided to select one of the objects. The order that the targets were displayed on each participants' screen was randomized to prevent the use of purely spatial cues (e.g. 'the one on the left'). To ensure that a single inattentive participant could not prevent the network from progressing, we included a forty-five second timer on each trial. If all dyads in the network responded within this time, they all immediately advanced to the next trial; if no response was recorded, they timed out and were automatically advanced. After a selection was made, both participants in a dyad were given full feedback and received bonus payment for correct responses.

After the communication phase, participants advanced to a post-test phase where they were asked to produce descriptions of the same tangrams for new participants to see in the future (Fig. 2C). Critically, we manipulated the target audience in a within-participant design. They were asked to provide descriptions both for new members of their *own* group and for new members of the *other* group. To control for possible order effects, we elicited these descriptions in two blocks corresponding to the two target audiences (*own* vs. *other*). We randomized both the sequence in which the four objects appeared within each block and the order of the two blocks.

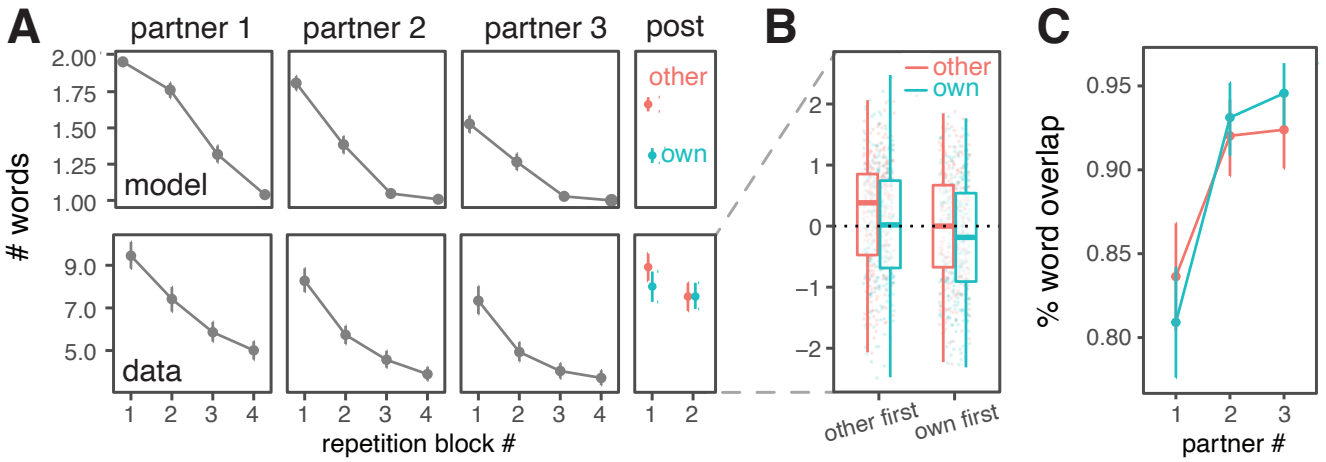

Figure 3: *Simulation and Experiment 1 results.* (A) Speakers are progressively more willing to extend efficient conventions to new in-group partners over time, but (B) are less willing to extend them to out-group members in the post-test phase (utterance length z-scored within participant for boxplot). (C) Descriptions produced for out-group members draw on more *content* from early trials while those produced for in-group members draw more on later trials. Error bars are bootstrapped 95% CIs.

Results

Generalization and partner-specificity within networks First, we were able to successfully replicate previous tests of generalization within a *single* community (see Hawkins et al., 2020). In a linear regression predicting (log) utterance length, we found that participants used shorter descriptions across multiple repetition blocks with the *same* partner, $b = -0.25$, $t(14) = -15$, $p < 0.001$; reverted to relatively longer descriptions $b = 0.5$, $t(36.1) = 9.6$, $p < 0.001$ on the first block with a new partner relative to the final block with a previous partner (consistent with partner-specificity); and were progressively more willing to use these shorter description on the initial trial with a new partner, $b = -0.16$, $t(81.7) = -4.4$, $p < 0.001$ (see Fig. 3A).

Speakers produce longer descriptions for out-group We now turn to the descriptions produced in the post-test phase. Our *community-sensitive* model predicts that participants expect conventions to be specific to the group context; thus, a convention that was only observed in the red group will only be expected to generalize to new red group members, not to blue group members. We again operationalized this prediction in terms of the length of the descriptions provided for each group, which can be considered a rough proxy for the amount of information the speaker believes they need to provide to a member of that group given their common ground (e.g. Fussell & Krauss, 1989). We ran a linear mixed-effects model predicting (log) number of words in each description including fixed effects of target audience (coded as *own* vs. *other*) and block order (coded as *first* or *second*) and the maximal random effects structure that converged: intercepts and main effects at the individual participant level, as well as intercepts for each network and for each target tangram. We found a significant main effect of target audience, with descriptions produced for a new member of the out-group ($m_1 = 8.9$ words) significantly longer than descriptions produced for a new member of the in-group ($m_2 = 8.0$ words), $t(166.7) = 3.24, p = 0.001$. For comparison, utterances produced for the *other* group were closest in length to the descriptions produced for one's very first partner in the communication phase (9.4 words). Meanwhile, utterances produced for an unseen member of one's *own* group were between the initial descriptions for one's second and third partners (8.3 and 7.3 words). Additionally, we found a significant order effect, with descriptions on the second block significantly shorter regardless of the target audience, $b = -0.14$, $t(170) = -4.07$, $p < 0.001$ (Fig. 3B), likely due to a combination of priming and eagerness to finish the experiment. There was no evidence supporting an interaction term, $\chi^2(5) = 0.92$, $p = 0.97$.

Out-group descriptions more similar to earlier trials Finally, we compare the *content* of post-test descriptions with the descriptions produced on earlier rounds, hypothesizing that out-group descriptions would avoid the group-specific conventions formed later in the game in favor of more generic features. We operationalized similarity using the set intersection of the pair of utterances: for every description of a target tangram produced by a speaker during the communication phase, we computed whether any words overlapped with their post-test utterance. We then conducted a mixed-effects logistic regression predicting the binary variable of whether utterances overlapped, including fixed effects of target audience (*own* vs. *other*) and time (partner 1, 2, or 3) as well as random intercepts for each tangram and speaker. First, we found an overall main effect of time with post-test descriptions more likely to contain words from later trials, $b = 0.77$, $z = 9.5$, $p < 0.001$. Critically, however, we also found an interaction with target audience, $b = -0.19$, $z = -2.4$, $p = 0.017$,

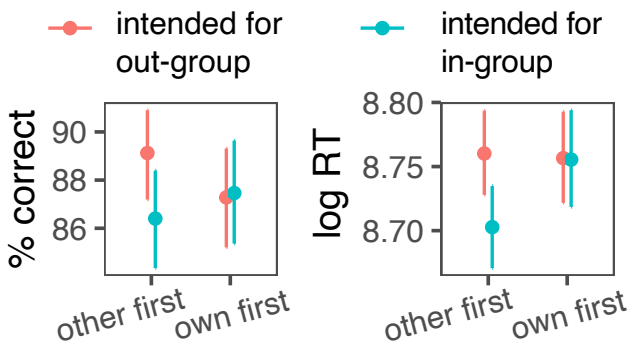

Figure 4: *Experiment 2 results.* Naive participants are marginally more accurate and slower to read the longer descriptions produced for out-group members than in-group members. Error bars are bootstrapped 95% CIs.

where *other*-intended utterances tended to overlap with the earliest utterances more than *own*-intended utterances and *own*-intended utterances tended to overlap more highly with the utterances produced with later partners (Fig. 3C). These results suggest that speakers not only produced longer descriptions for out-group members but also tailored the lexical content of their descriptions accordingly.

Experiment 2: Evaluating transparency

Our first experiment suggested that speakers distinguish conventions that are likely to be meaningful only within their own group from those that are likely to be shared more universally. While we measured this distinction in terms of different description lengths, it is unclear whether these differences in length actually reflect differences in the *transparency* of the convention (Atkinson, Mills, & Smith, 2019). That is, are the shorter descriptions produced for one's own group actually more difficult for a naive listener to understand? In this experiment, we empirically evaluate the effect of intended audience on downstream comprehension. We showed the descriptions elicited in the post-test of Experiment 1 to an independent sample of naive participants and asked them to select which tangram object was being described. We predicted that messages intended for in-group members may rely to a greater extent on common ground derived from their idiosyncratic interactions with other group members, and therefore lead to lower recognition accuracy than messages intended for out-group members.

Methods

We recruited 500 fluent English speakers from Prolific to complete a short survey implemented with jsPsych (De Leeuw, 2015). We excluded participants that failed a catch trial ("click the one that's furthest to the left") as well as any trials with (log) response times outside two standard deviations of the group mean (i.e. $<$ 1s or $>$20s) which could indicate a lapse in attention. The 1,056 descriptions produced in the post-test of Experiment 1 were partitioned into subsets corresponding to each of the 8 tangrams. We showed participants exactly one description of each tangram, sampled randomly from these sub-sets and presented in a randomized order, for a total of 8 trials. Each trial proceeded as a simple recognition task: participants were shown a description and the corresponding set of four tangrams (A or B) from the community where it was produced. Participants were then asked to click on the tangram that best matched the description and rated their confidence using a slider ranging from 'not at all confident' to 'very confident.' To prevent learning within this short task, participants were not given feedback on their responses. Finally, to ensure that we received a sufficient number of responses for each description under our randomization scheme, we ran the experiment in several large batches, removing descriptions from the candidate set once they appeared more than five times. Although this procedure necessarily created an imbalanced sample, we ensured that all descriptions were seen at least once; the modal responses per description was three.

Results

Comprehension accuracy Our primary prediction concerned the relative transparency of descriptions produced for in-group members vs. out-group members. To test this prediction, we ran a logistic mixed-effects model predicting the binary correctness of each response. We included fixed effects of the description's original intended audience (coded *own* vs. *other*) and the order that the speaker produced them in (coded as *own-first* vs. *other-first*), as well as their interaction. We included the maximal random effects structure that converged, with random intercepts for each original speaker, for each original network the speaker belonged to, and each tangram item. We found a weak simple effect of target audience for descriptions in the *other-first* group, $b = -0.27,\ z = -2.01, p = 0.044$ as well as marginal evidence suggesting an interaction, $b = 0.29, z = 1.52, p = 0.12$, clarifying that there appeared to be no such audience effect for descriptions produced in the reverse order (see Fig. 4A).

Response time We also considered an analogous analysis for (log) response times recorded in our comprehension task. Response time is clearly confounded with description length (i.e. longer descriptions take longer to read before a response can be made), and we found results consistent with description length differences observed in Experiment 1. Participants responded more slowly for longer, *other*-intended utterances than shorter *own*-intended utterances in the other-first group, $b = -0.06,\ t(85.7) = -2.7,\ p = 0.008$, and a weak interaction suggests that this effect was limited to the *own*-first order, $b = 0.07, t(95.1) = 2.3, p = 0.02$ (see Fig. 4B). These results are consistent with a weak but reliable difference in the external transparency of descriptions originally produced for in-group and out-group members, although they were similarly affected by order effects.

General Discussion

How do speakers know which conventions to use for different partners? In this paper, we argued that the ability to code-switch requires common ground to be represented not only at the level of specific partners, but also to be sensitive to the *communities* those partners belong to. We formalized this idea by extending a recent hierarchical Bayesian model of convention formation with an intermediate layer representing latent group membership and tested the predictions of this model in two behavioral experiments implementing a minimal group paradigm with small networks of interacting participants. Even under these weakly induced and short-lived groups, we found that speakers were sensitive to the group membership of their audience, producing marginally shorter and more transparent utterances for out-group members.

While these findings support the qualitative predictions of our model, effect sizes in both experiments were smaller than expected. Several contributing factors are possible. First, our minimal group manipulation may not have been convincing. Groups were based on arbitrary color assignments, indicated only by the avatars of different partners, which may not have been sufficiently salient to mark differences between groups. Even for participants that attended to distinctions between in-group and out-group members, it is possible that they were legitimately not convinced that in-group conventions would generalize to a hypothetical future member of their group. Different participants may have made different assumptions about this hypothetical person that limit generalizability relative to what would be expected in a fourth block of interaction with real partners. Indeed, the primary discrepancy from our model predictions was an insufficient *decrease* in description length for new in-group members rather than a failure to limit extensions to out-group members. Second, it is likely that not all decreases in utterance length reflect substantive decreases in transparency. Different groups may converge on different conventions while both conventions remain understandable to naive observers. Hence, the size of the transparency effect may be particularly small and requires further confirmatory replication.

Our work raises several key open questions. First, a direct corollary of our model is that speakers with well-calibrated representations of the language used by different social groups should intentionally choose maximally diagnostic in-group conventions to signal their own identity. Indeed, we predict that this signaling behavior is directly related to the value of group membership in the agent's environment: for example, lower-status individuals are more likely to over-use jargon and competitive settings tend to increase signalling behavior. Second, it is unclear how this framework ought to extend beyond social conventions to prescriptive or moral norms. Even for young children, the latter may be expected to generalize more universally across groups (Schmidt, Rakoczy, & Tomasello, 2012). This behavior may simply reflect assumptions about the variance at different layers of the hierarchy, or may require a different generative model entirely. Finally, while we assumed for simplicity that each individual belongs to a single group, it is important to extend our model to the case of multiple overlapping groups that vary in their status. Indeed, code-switching not only allows an individual to adjust their language for in-group and out-group members, but also allows them to pass as a credible member of multiple groups. More broadly, we hope that incorporating explicit representations of social group identity into cognitive models of communication and convention may open pathways to better capturing the diversity of experiences and linguistic identities within the broader language community and the challenges that accompany inter-group communication.

Acknowledgements

Thanks to Kenny Smith, Olga Feher, and Herb Clark for helpful discussions. This work was supported by NSF grant #1911835 to RDH, AEG, and TDG.

Materials and code for reproducing all experiments, analyses, and model simulations available at:
`https://github.com/hawkrobe/code_switching`

References

Almaatouq, A., Becker, J., Houghton, J. P., Paton, N., Watts, D. J., & Whiting, M. E. (2021). Empirica: a virtual lab for high-throughput macro-level experiments. *Behavior Research Methods*, 1–14.

Atkinson, M., Mills, G. J., & Smith, K. (2019). Social group effects on the emergence of communicative conventions and language complexity. *Journal of Language Evolution*, *4*(1), 1–18.

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

Brown-Schmidt, S., Yoon, S. O., & Ryskin, R. A. (2015). People as contexts in conversation. In *Psychology of learning and motivation* (Vol. 62, pp. 59–99). Elsevier.

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

DeBose, C. E. (1992). Codeswitching: Black English and standard English in the African-American linguistic repertoire. *Journal of Multilingual & Multicultural Development*, *13*(1-2), 157–167.

De Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior research methods*, *47*(1), 1–12.

Eble, C. C. (1996). *Slang & sociability: In-group language among college students*. Chapel Hill, NC: UNC Press.

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.

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.

Gardner-Chloros, P. (2009). *Code-switching*. Cambridge, UK: Cambridge University Press.

Gershman, S. J., Pouncy, H. T., & 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.

Gumperz, J. J. (1982). *Discourse strategies*. Cambridge, UK: Cambridge University Press.

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 42nd Annual Meeting of the Cognitive Science Society.*

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

Katzner, K., & Miller, K. (2002). *The languages of the world*. Routledge.

Kerr, D., & Smith, K. (2016). The spontaneous emergence of linguistic diversity in an artificial language. In *EVOLANG.*

Lau, T., Pouncy, H. T., Gershman, S. J., & Cikara, M. (2018). Discovering social groups via latent structure learning. *Journal of Experimental Psychology: General*, *147*(12), 1881.

Partridge, E. (2006). *A dictionary of slang and unconventional english*. London: Routledge.

Schmidt, M. F., Rakoczy, H., & Tomasello, M. (2012). Young children enforce social norms selectively depending on the violator's group affiliation. *Cognition*, *124*(3), 325–333.

Tajfel, H. (1982). Social psychology of intergroup relations. *Annual Review of Psychology*, *33*(1), 1–39.

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.