Précis of *Coordinating on meaning in communication*

Robert D. Hawkins
Department of Psychology, Stanford University

It is tempting to think of word meanings as entries in a dictionary shared by speakers of a language. Drop into any conversation between friends, however, and you wade into a stream of shorthand, jargon, slang, references, nicknames, and inside jokes — some of which you may understand, but the rest of which may be meaningful to them alone. There is no guarantee, it seems, that any two speakers of a language will share the same dictionary. To make matters worse, we live in an ever-changing world where we need to talk about new things. There is also no guarantee that any dictionary could anticipate the meanings we will need to express in each new context. If we cannot rely on the existence of a shared dictionary for coordination, what can we rely on? How do we manage to understand one another so effortlessly in this patchwork landscape of meaning?

My dissertation investigates the cognitive mechanisms that allow individuals, and communities, to solve this challenge. The core theoretical contribution is an account of communication relaxing the assumption that speakers of a language share the same "protocol." Instead, we propose that communication is better understood as a multi-agent meta-learning problem guided by social inference. Agents must integrate background expectations about reliable community-wide conventions with new *ad hoc*, partner-specific pacts constructed on the fly. Chapters 2 formalizes this proposal in a hierarchical Bayesian model and presents simulation results capturing key effects from the literature. Chapter 3 evaluates an algorithm scaling this proposal to a recurrent neural network that interactively adapts to human partners using natural language. Chapter 4 introduces a large behavioral corpus that provides a higher resolution look at the dynamics of *ad hoc* convention formation. Chapter 5 tests how the communicative needs of the context shape the resulting *ad hoc* conventions. Finally, Chapter 6 assesses the generality of the proposed mechanisms by examining convention formation in a *graphical* Pictionary task.

Understanding the role of learning and social cognition in everyday language use is a foundational question at the intersection of psychology, neuroscience, and linguistics. The methodological combination of computational modeling and naturalistic communication experiments gives new insight into how such rich, flexible language behavior may arise from these general-purpose cognitive mechanisms. The proposed theory builds on accounts of social convention and meaning developed in analytic philosophy, particularly by Lewis (1969) and Davidson (1986), and suggests broader application potential toward building more adaptive and socially intelligent AI, consistent with current directions in computer science. More broadly, this work opens avenues for future work to understand how social conventions and norms are represented in the mind, how these conventions are shaped by local context, and how they give rise to the remarkable feats of social coordination that are so distinctive of human cultures (Hawkins, Goodman, & Goldstone, 2019).

Introduction

A core function of communication is *reference*: using words to convey the identity of an object in the environment (Brown, 1958; Searle, 1969; van Deemter, 2016). Throughout my dissertation, I use a rich, naturalistic communication paradigm called a *repeated reference game* (Krauss & Weinheimer, 1964) that requires a pair of participants to find some way of referring to novel, ambiguous stimuli they don't already have strong conventions for. On each trial, one participant (the director) is privately shown a *target* object in an array of distractors and must produce a referring expression allowing their partner (the matcher) to correctly select that object from the array. Critically, each object appears as the target multiple times in the trial sequence, allowing the experimenter to examine how referring expressions change as the director and matcher accumulate a shared history of interaction, or 'common ground' (Clark, 1996). To the extent that the director and matcher converge on a stable and accurate system of referring expressions, and these referring expressions differ from the ones that were initially produced, it may be claimed that *ad hoc conventions* or *pacts* have been constructed within the dyad.

One of the earliest and most intriguing phenomena observed in this task is that descriptions are dramatically shortened across repetitions: an initial description like "the one that looks like an upside-down martini glass in a wire stand" may gradually converge to "martini" by the end. That is, speakers are able to communicate the same referential content much more efficiently over time. Subsequent work has established a number of signature properties of this process through careful experimental manipulation. First, the extent to which descriptions are shortened is *contingent* on evidence of understanding from the matcher (Krauss & Weinheimer, 1966; Krauss et al., 1977; Hupet & Chantraine, 1992), and is therefore not easily explained as a mere practice or repetition effect. Second, the resulting labels are *partner-specific* in the sense that they do not transfer if a novel matcher is introduced (Wilkes-Gibbs & Clark, 1992; Metzing & Brennan, 2003; Brennan & Hanna, 2009). Third, they are *sticky* in the sense that they persist through precedent with the same partner even after the referential context changes (Brennan & Clark, 1996), and are readily extended to similar objects (Markman & Makin, 1998).

These qualitative effects provide a core empirical backbone for theories of communication to explain, but immediately pose difficulties for current computational models. One of the most promising candidates is a family of probabilistic models which derive flexibility in meaning from *de novo* Gricean reasoning, formalized as recursive social inference (Goodman & Frank, 2016; Franke & Jäger, 2016). These models show how listeners use context to draw enriched inferences about intended meaning even in cases of ambiguous or non-literal usage (Clark, 1983; Lascarides & Copestake, 1998; Glucksberg & McGlone, 2001). While such one-shot pragmatic reasoning has provided a powerful explanation of many forms of flexibility, it remains insufficient to explain the dynamics of how new meanings become shared *over time*. Like other models of communication descended from Shannon (1948), the problem is that recursive reasoning still bottoms out in a fixed 'dictionary' of literal meanings assumed to be already shared between speakers. To overcome this problem, I turned to the literature on language acquisition and language evolution, which focus on how meaning arises in the first place. I suggest that the same learning mechanisms may play a ubiquitous role in establishing meaning even in ordinary interactions between adults.

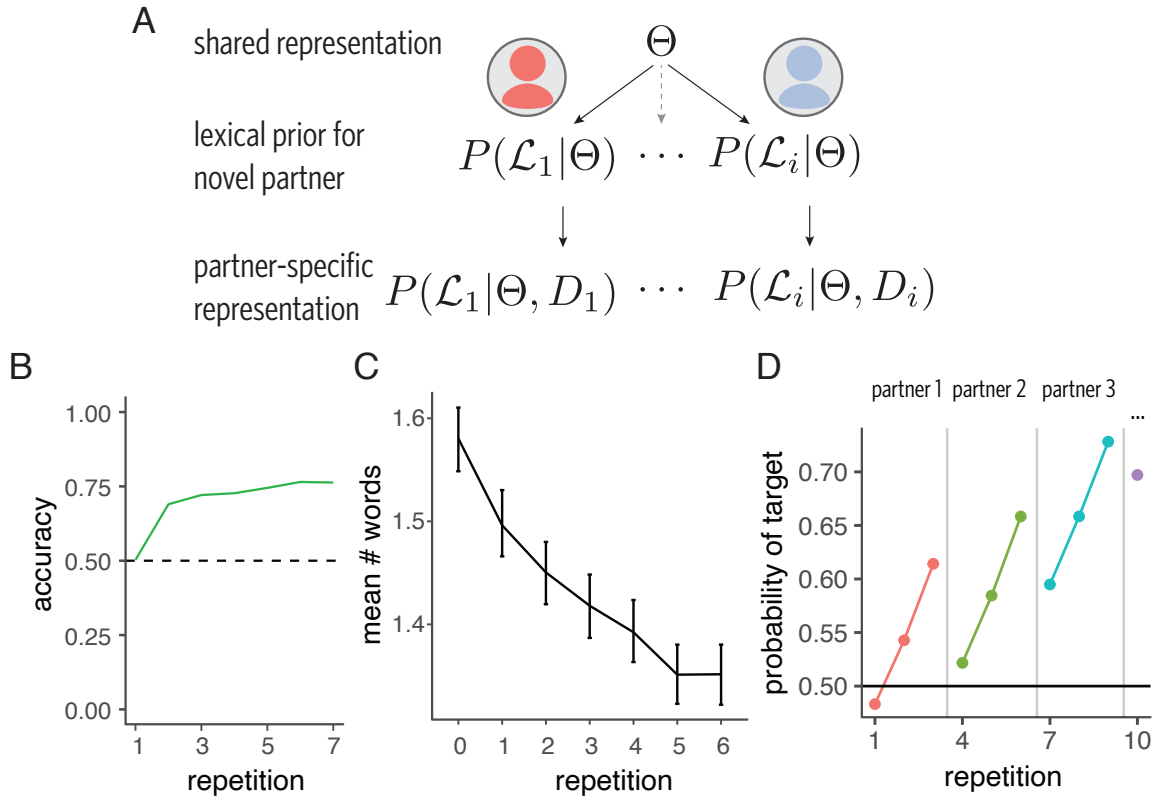

Figure 1: Model and simulation results from Chapter 2. (A) Schematic of hierarchical Bayesian model. (B) Accuracy rises as speaker and listener coordinate. (C) Utterances reduce as speaker becomes more confident of partner's meaning. (D) Model initially represents meanings as partner-specific but generalizes convention after several partners use same meanings.

Chapter 2: An inferential model of convention formation

A key difficulty of rejecting the assumption of a shared 'dictionary' is that adults *do* nonetheless bring strong background expectations about meanings into their interactions. After all, these expectations are precisely the end-point that models of language acquisition and language evolution seek to explain, over different timescales. How can agents both maintain stable beliefs about what will be meaningful to new partners, and also flexibly deviate from these meanings on a partner-specific basis? This apparent tension between stability and flexibility suggests three functional desiderata for computational models of meaning:

1. a mechanism for how community-wide *conventions* (Lewis, 1969) shape initial interactions with a new partner.
2. a mechanism for how new *ad hoc* conventions can be so rapidly constructed with an individual partner.
3. a mechanism connecting the two: how and when are meanings generalized from individuals to communities?

In **Chapter 2**, I propose a hierarchical Bayesian model (HBM) of convention formation and argue that it satisfies these three desiderata. HBMs have been key to understanding how the human mind solves a range of other difficult inductive problems where abstract, shared properties must be jointly inferred with idiosyncratic particulars of instances, including causal learning (Kemp, Goodman, & Tenenbaum, 2010; Goodman, Ullman, & Tenenbaum, 2011), speech perception (Kleinschmidt & Jaeger, 2015) and concept learning (Kemp, Perfors, & Tenenbaum, 2007). In the context of communication, the first step is to replace the shared 'dictionary' of meanings with a distribution representing "lexical uncertainty" over different possible dictionaries a particular partner might be using (see Cooper, Dobnik, Larsson, & Lappin, 2015; Bergen, Levy, & Goodman, 2016; Smith, Goodman, & Frank, 2013; Potts, Lassiter, Levy, & Frank, 2016). In an HBM, this uncertainty is assumed to be hierarchically structured (see Fig. 1A). The top level is a stable prior representing abstract conventions shared across all agents in the community, while idiosyncratic partner-specific knowledge is represented at the lower level. This roughly corresponds to the distinction made by Davidson (1986) between a "prior theory" and a "passing theory." In other words, conventions supply a useful but often uncertain first guess about the linguistic meanings that will be shared with a new partner.

Critically, these expectations may then be updated using evidence of a partner's language use. Specifically, agents use Bayesian theory of mind to make inferences about that partner's latent 'dictionary,' leading to local *ad hoc* conventions. As each party learns about the other and attempts to be understood under their updated model of partner-specific meaning, they may coordinate on a system of meaning that is tailored to be efficient and accurate for their present purposes. Thus the shared semantic prototype can be thought of as a backbone supporting rapid learning for new partners and situations, and may in turn be learned by pooling across many extended interactions with individuals. In summary, this model formalizes a functional view of linguistic conventions as solutions to a meta-learning problem, in which every agent is simultaneously seeking to infer and use the conventions that other agents are using, to support more successful communication.

I evaluated this model through a series of simulations. First, I showed that two agents updating their beliefs about meaning in this way can coordinate even in the absence of strong initial priors (Fig. 1B). Their initial choices can be taken as evidence for a particular lexicon and become the basis for successful communication. Second, I showed that a preference for less costly utterances combined with learning gives rise to shorter descriptions over time (Fig. 1C). The speaker model initially produces longer utterances as a way of hedging against their own uncertainty about what meanings will be understood, but omits words over time as they become more confident. Third, I tested the model's pattern of generalization over different partners (Fig. 1D). When a new partner is first introduced, it reverts nearly to its original state, displaying the signature phenomenon of partner-specificity. As it interacts with different partners using language in a similar way, however, these *ad hoc* conventions gradually become bona fide *community-level* conventions that are readily extended to initial encounters with new partners. These simulations of key qualitative phenomena suggest that the proposed HBM is a promising computational account of how conventions are represented and learned at different scales.

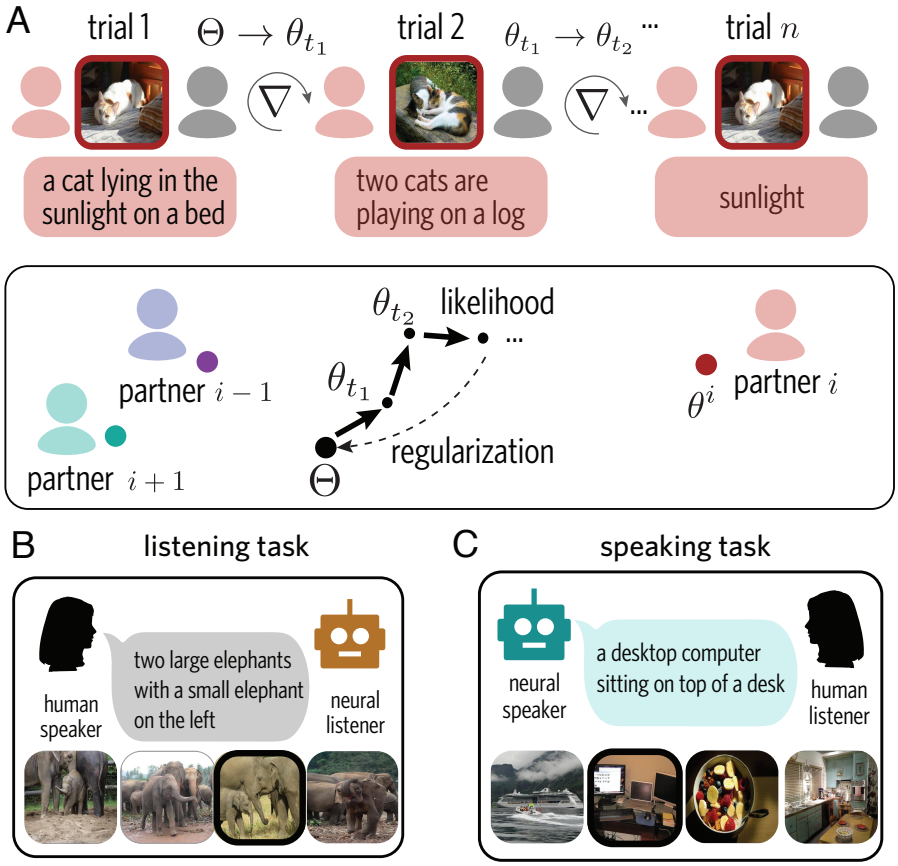

Figure 2: Modeling approach and experiments from Chapter 3. (A) Schematic of a regularized continual learning approach that allows artificial agents to rapidly adapt to their partner through interaction, which was evaluated using both (B) an interactive listening task where the agent interpreted referring expressions from a human speaker and (C) a speaker task where the agent produced descriptions for a human listener.

Chapter 3: Coordination between humans and machines

While the hierarchical Bayesian model explored in Chapter 2 is shown to have desirable theoretical properties, it also becomes intractable to use as the vocabulary and object set grows, making it untenable for modeling arbitrary natural language. In **Chapter 3**, I instead take the community-level expectations to be an initialization for the weights of an image-captioning neural network. Just as probabilistic models have assumed a fixed 'dictionary,' approaches based on deep neural networks typically learn a monolithic meaning function during training, with fixed weights during use. I exploit theoretical connections between the hierarchical Bayesian framework and recent approaches to multi-task- and meta-learning in deep neural networks (Nagabandi, Finn, & Levine, 2018; Grant, Finn, Levine, Darrell, & Griffiths, 2018; Jerfel, Grant, Griffiths, & Heller, 2018) to propose an online continual learning approach for a neural image-captioning model. Specifically, the community-level prior corresponds to a pre-trained, task-general initialization. Condition-

ing on new data from a particular partner corresponds to regularized gradient descent (Fig. 2A).

To evaluate this approach, I implemented a repeated reference game using images from the validation set of COCO (Lin et al., 2014) as the targets of reference. I considered the model's ability to form *ad hoc* conventions with human partners in two different tasks, which presented distinct challenges for its initialized background expectations. First, in a *listening* task (Fig. 2B), the model is paired with a human speaker and must learn to interpret their natural referring expressions in challenging contexts. Chatbox messages were sent to a GPU where the model weights from the previous trial were loaded, used to generate a response, and updated in real-time for the next round with a latency of approximately 6-10 seconds. The listener model initially performs much less accurately than a baseline human listener because of the idiosyncratic and out-of-sample language produced by the human speaker. Yet our model rapidly improves in accuracy as it fine-tunes a parthner-specific representation. Second, in a *speaking* task (Fig. 2C), the model is paired with a human listener and must learn to *generate* appropriate referring expressions in easier contexts. The model initially produced more complex referring expressions than required to distinguish the images, due to biases from its training corpus, but it became dramatically more efficient over the course of the interaction while maintaining high accuracy. These results validate the proposed inferential model of convention formation on real natural-language input and output, and suggest further avenues of application potential for adaptive artificial intelligence based on basic cognitive principles.

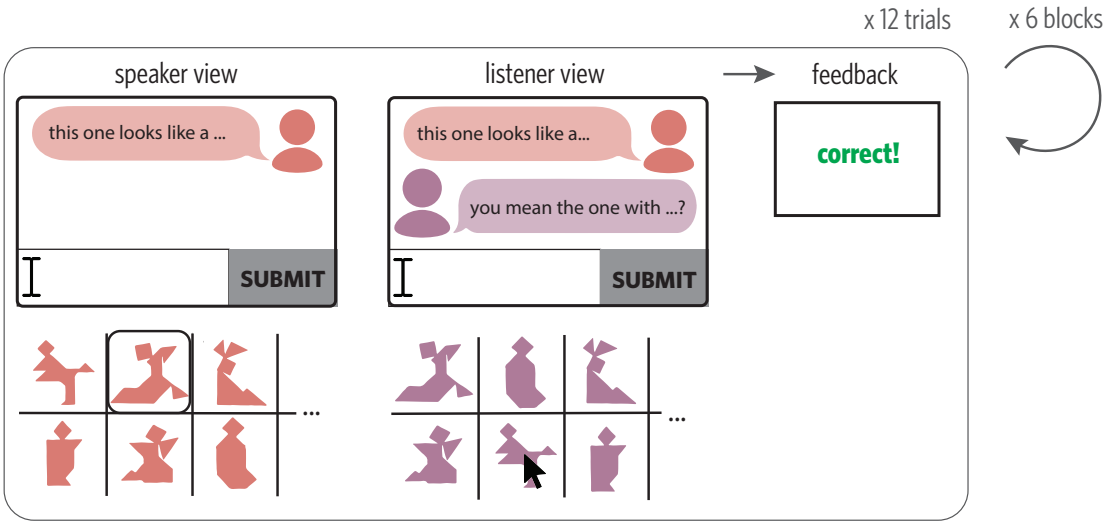

Figure 3: Experimental design for repeated reference game used in Chapter 4. Participants were paired over the web and communicated freely using a chat box. After a selection is made, the director is given feedback about which object was selected, and the matcher is given feedback about the true target object. Tangram stimuli were reproduced from Clark & Wilkes-Gibbs (1986).

Chapter 4: Characterizing the dynamics of coordination

As theories are increasingly formalized as computational models like those in the previous chapters, finding criteria to distinguish between them will depend critically upon resolving more detailed theoretical questions about the dynamics of adaptation in natural language communication. **Chapter 4** describes a new, open corpus ($> 15,000$ messages) of repeated reference games (see Fig. 3) and a variety of analyses that address current gaps in measurement and establish a firmer theoretical foundation for future modeling work. This effort addresses two methodological challenges that have limited the ability of previous studies to provide a sufficiently fine-grained characterization of behavior. First, more data was needed. Recent technical developments have allowed interactive multi-player experiments to be run on the web (Hawkins, 2015), boosting sample sizes by an order of magnitude. Second, the computational techniques needed to work with rich natural language data were limited at the time of prior work, but have become newly tractable given developments in natural language processing (NLP).

The analyses in this chapter roughly divide into two categories, corresponding to the dynamics of syntactic *structure* and semantic *content*. The investigations of syntactic structure focus on the process by which referring expressions are gradually shortened to communicate the same idea more efficiently. One particularly simple model, for example, might predict that shortening is purely driven by a random corruption process: at each repetition, each word from the previous repetition's utterance has some probability of being dropped. Raw word counts alone are not sufficient for disambiguating this simple model from more cognitively complex proposals. To move beyond word counts, I extracted part-of-speech tags and syntax trees from the text to understand which parts of utterances were being dropped, and in which sequence. In contrast to the predictions of the random corruption model, I find that clauses and modifiers tend to be dropped in clusters, preferentially leaving open-class parts of speech (e.g. an adjective and noun) by the final repetition, and that the choice to shorten an utterance or not depends on sources of listener feedback.

Next, I examine the semantic content of utterances over the course of this shortening process. These analyses revolve around the theoretical constructs of *arbitrariness* and *stability*, which have been central to accounts of convention since Lewis (1969). Arbitrariness refers to the claim that multiple equally successful solutions exist in the space of possible conventions: there is no single optimal solution that all speakers should objectively use. Stability refers to the claim that, once a solution has been found, speakers should not deviate from it. I operationalize these claims in the high-dimensional space of vector embeddings for referring expressions (i.e. GloVe embeddings). By measuring the similarity between referring expressions in this space, I find that signatures of arbitrariness and stability gradually increase over the course of the interaction. Additionally, while different pairs coordinate on a wide range of idiosyncratic solutions to the problem of reference, they do so in a highly path-dependent manner: words that are more discriminative in the initial context (i.e. that were used for one target more than others) are more likely to persist through the final round. Taken together, these findings characterize core processes operating within the microcosm of dyadic, natural-language interactions.

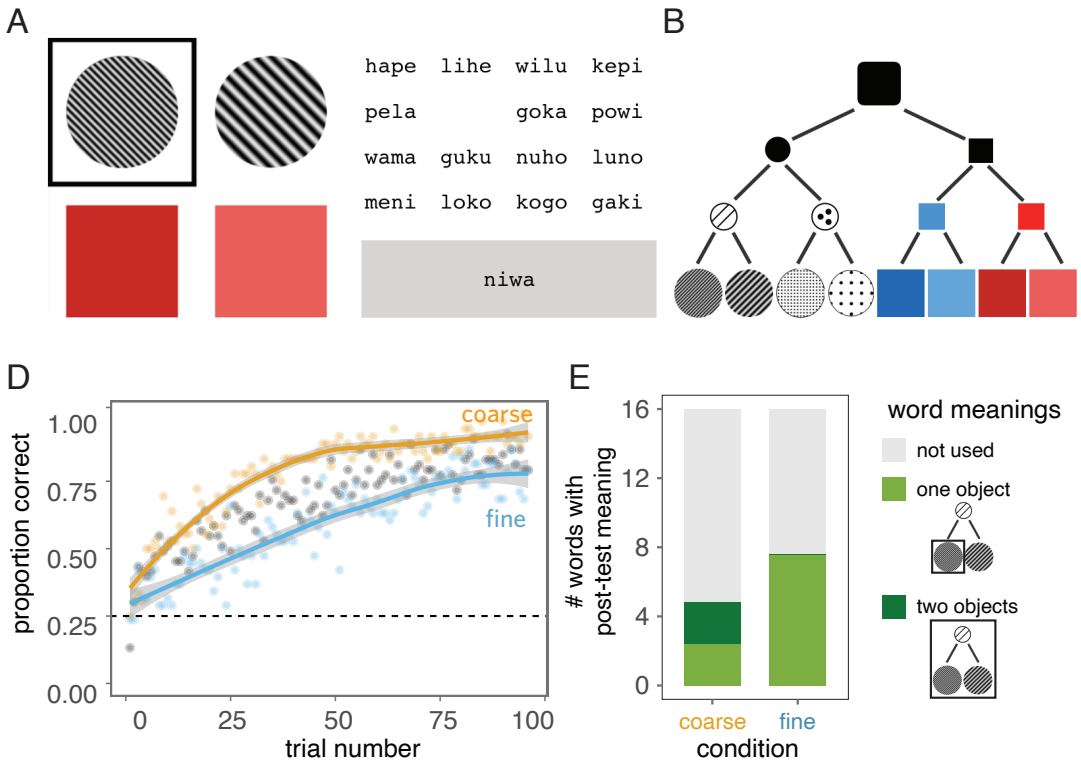

Figure 4: Methods and results from Chapter 5. (A) Example of *fine* context where one of the distractors belongs to the same fine-grained branch of the hierarchy as the target (i.e. another striped circle), so any abstract label would be insufficient to disambiguate them. The target is highlighted for the speaker with a black square. (B) Drag-and-drop chat box interface with artificial labels. (C) Hierarchical organization of stimuli. (D) Participants became increasingly accurate over repeated interaction. (E) Statistics of environment lead to different systems of meaning: participants developed more abstract meanings (applying to 2 objects) than specific meanings (applying to 1 object) in the coarse condition.

Chapter 5: How does communicative context shape conventions?

The rapid timescale of adaptation observed in dyadic interaction is not only of interest for cognitive theories of meaning and social coordination. It may also be a key building block toward understanding the remarkable adaptiveness and efficiency that is characteristic of human languages over longer time scales. Recent computational approaches to language evolution have argued that the lexical conventions of languages balance simplicity, or learnability, with usage needs (Winters, Kirby, & Smith, 2014; Regier, Kemp, & Kay, 2015; Kirby, Tamariz, Cornish, & Smith, 2015; Gibson et al., 2019). For example, languages in warm regions ought to be more likely to collapse the distinction between ice and snow into a single word, simply because there are fewer occasions that require distinguishing between the two (Regier et al., 2016). While this is powerful explanatory principle, prior work has largely focused on the functional *outcomes* of the overall language evolution process rather than the individual-level cognitive mechanisms driving the process itself. If community-wide

conventions emerge from agents generalizing across dyadic interactions, then understanding the local mechanisms leading to efficiency and informativity *within* a dyad may help explain how a community's conventions become well-calibrated to their environment.

Chapter 5 examines how the cognitive mechanisms proposed in earlier chapters allow context to shape *ad hoc* conventions (Hawkins, Franke, Smith, & Goodman, 2018). Specifically, I propose that context-sensitivity results from the impact of inductive biases on meaning inference, specifically the Gricean expectations that a partner is trying to be appropriately informative. To isolate these mechanisms from the background priors speakers bring into new interactions, I adopted an artificial-language paradigm from work on language evolution. Participants were paired to communicate about a set of target objects (Fig. 4A) through a chatbox containing a fixed vocabulary (Fig. 4B) and given full feedback after each trial. The stimuli were designed to visually cluster in a three-level hierarchy (Fig. 4C). Critically, the statistics of the environment were manipulated across different pairs to make different distinctions relevant. In the 'fine' condition, every context contained distractors that were very similar to the target, thus inducing a communicative need for finer distinctions. In the 'coarse' condition, distractors always differed from the target at higher levels of the concept hierarchy. I hypothesized that pairs should coordinate on distinct names for each object when the context frequently makes fine distinctions relevant. Conversely, they should converge on a more efficient and compressed system of conventions for abstract categories in coarser contexts, even if a finer mapping would be sufficient.

Over 100 trials, dyads were able to successfully coordinate on reliable systems (Fig. 4D). Using a post-test probing the resulting conventions and a statistical approach inferring participants' beliefs about conventions at earlier stages of the task, I found systematic differences in the *level of abstraction* of meanings that formed across these different conditions. In the coarse condition, participants learned to communicate with smaller vocabularies collapsing over distinctions within categories, while in the fine condition they coordinated on specific labels for sub-categories (Fig. 4E). These differences arose even though both groups referred to each target object the same number of times; the only difference was the identity of distractors. These results indicate that lexical conventions are shaped by pragmatic considerations of *informativity in context* as each speaker attempts to infer what the other intends to mean. Our minds organize the world into meaningful conceptual hierarchies but our shared language may only come to reflect this structure when it is communicatively relevant.

Chapter 6: How general are mechanisms across modalities?

While linguistic communication is powerful and prevalent, research on the dynamics of adaptation in other communication modalities, including drawing (Garrod, Fay, Lee, Oberlander, & MacLeod, 2007; Galantucci, 2005; Healey, Swoboda, Umata, & King, 2007; Theisen, Oberlander, & Kirby, 2010; Fay, Garrod, Roberts, & Swoboda, 2010) and gesture (Goldin-Meadow, McNeill, & Singleton, 1996; Goldin-Meadow & McNeill, 1999), is important for two reasons. First, a core claim of the proposed hierarchical Bayesian model is that the learning and social cognition mechanisms underlying communication are domain-general. In other words, there is nothing special about spoken or written language; any ad hoc system that we use to send messages should display similar learning dynamics. In all cases we are

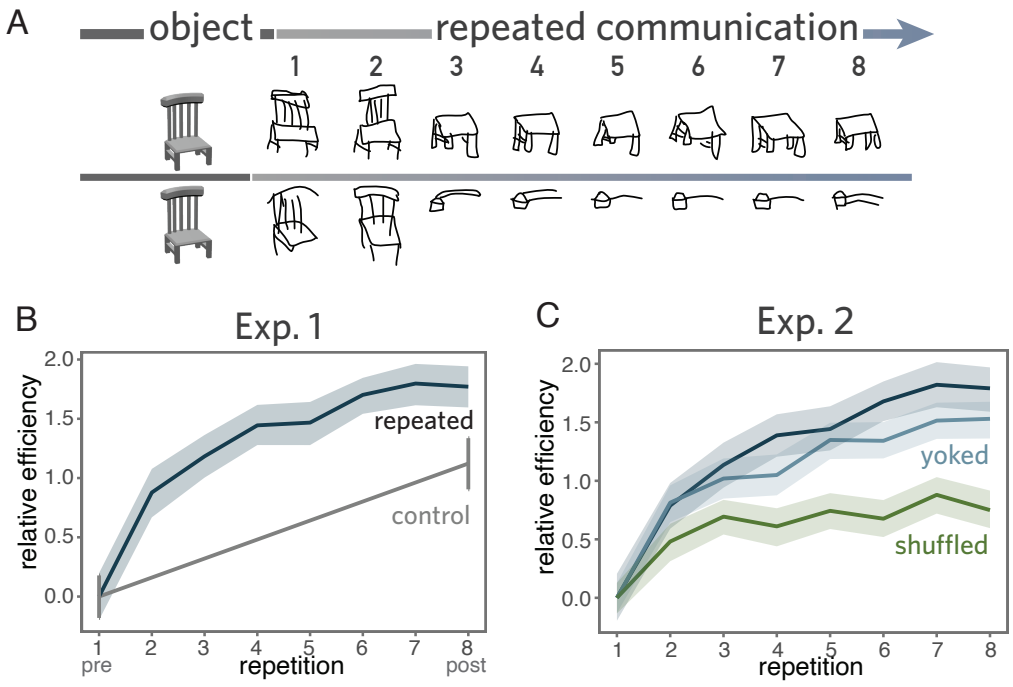

Figure 5: Results from the graphical reference game in Chapter 6. (A) Examples of two dyads depicting the same object over successive trials (B) Performance improves for repeated objects more than for control objects only seen at the beginning and end. (C) Drawings were harder for naive viewers to recognize in the absence of interaction history.

trying to coordinate on meaning with other minds. Second, because this model also claims a critical role for background priors about what is expected to be meaningful, different communication modalities should nevertheless display certain systematic differences. For example, the abstract tangram shapes used in Chapter 4 were hard to name but would be easy to draw. Conversely, common objects like pieces of furniture are easy to refer to in the linguistic modality (e.g. 'armchair', 'lawn chair') due to strong background conventions, but hand-drawing the necessary distinctions may be costly, creating a functional need to form local conventions.

In **Chapter 6**, I explored the hypothesis that extended visual communication may promote the development of increasingly idiosyncratic yet effective ways of depicting objects (Hawkins, Sano, Goodman, & Fan, 2019). Specifically, I designed an interactive drawing-based reference game (i.e. *Pictionary*) similar to those used in previous chapters in which two participants repeatedly referred to targets in a context of chair images. I examined both how their task performance and the drawings they produced changed over time (see Fig. 5A). There are three aspects of this current work that advance previous work in this modality (e.g. Garrod et al., 2007). First, I included a set of control objects that were not repeatedly drawn but only shown at the beginning and end of the interaction, allowing measurement of the specific contribution of repeated reference vs. general practice effects. These control objects did not show the same performance gains as the repeated objects

(Fig. 5B), suggesting that conventions were object-specific. In a second experiment, I measured how strongly the visual properties of drawings drive recognition in the absence of interaction history for naive viewers (the *scrambled* condition), while equating other task variables (the *yoked* condition). Drawings became substantially harder for independent viewers to recognize without sharing the full interaction history (Fig. 5C). *Third*, I employed a convolutional neural network to quantitatively characterize changes in the high-level visual properties of drawings across repetitions (Simonyan & Zisserman, 2014; Fan, Yamins, & Turk-Browne, 2018)—a graphical analog of the semantic embeddings used in Chapter 4. Drawings became increasingly consistent within an interaction, but different pairs discovered different equilibria in the space of viable graphical conventions. The strong correspondence between these results and those found using linguistic communication suggests that the same generic mechanisms may support coordination across communication modalities.

Conclusions

Linguistic communication is one of the most difficult coordination problems in human social life. We use words to convey everything from the complex aroma of a wine to the esoteric constructs of a scientific theory. But language is not a monolithic body of knowledge; there is no complete and universal dictionary we can rely on to provide us the same meanings for the same words. The goal of my dissertation was to understand the cognitive mechanisms that nonetheless allow us to understand each other. Toward this end, I proposed a computational model of convention formation that reverse-engineers the flexibility of linguistic meaning from rational principles of hierarchical inference and pragmatic reasoning. In conjunction with finer-grained evidence from naturalistic social interactions, this theoretical approach provides a unified view of classic psycholinguistic phenomena, grounds community-level conventions in individual and dyadic processes, and lays the groundwork for more socially adaptive AI. By computationally characterizing these cognitive mechanisms, we will be better positioned to understand how interactions between individuals collectively give rise to human culture.

References

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

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

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

Brown, R. (1958). How shall a thing be called? *Psychological Review*, *65*(1), 14.

Clark, H. H. (1983). Making sense of nonce sense. *The process of language understanding*, 297–331.

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

Cooper, R., Dobnik, S., Larsson, S., & Lappin, S. (2015). Probabilistic type theory and natural language semantics. *LiLT (Linguistic Issues in Language Technology)*, *10*.

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

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

Fay, N., Garrod, S., Roberts, L., & Swoboda, N. (2010). The interactive evolution of human communication systems. *Cognitive Science*, *34*(3), 351–386.

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

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

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

Gibson, E., Futrell, R., Piandadosi, S. T., Dautriche, I., Mahowald, K., Bergen, L., & Levy, R. (2019). How efficiency shapes human language. *Trends in cognitive sciences*.

Glucksberg, S., & McGlone, M. S. (2001). *Understanding figurative language: From metaphor to idioms* (No. 36). Oxford University Press on Demand.

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

Goldin-Meadow, S., McNeill, D., & Singleton, J. (1996). Silence is liberating: removing the handcuffs on grammatical expression in the manual modality. *Psychological Review*, *103*(1), 34.

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

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

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

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

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

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

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

Jerfel, G., Grant, E., Griffiths, T. L., & Heller, K. (2018). Online gradient-based mixtures for transfer modulation in meta-learning. *arXiv:1812.06080*.

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

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

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

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

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

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

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

Lascarides, A., & Copestake, A. (1998). Pragmatics and word meaning. *Journal of linguistics*, *34*(2), 387–414.

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

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *European conference on computer vision* (pp. 740–755).

Markman, A. B., & Makin, V. S. (1998). Referential communication and category acquisition. *Journal of Experimental Psychology: General*, *127*(4), 331.

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), 201–213.

Nagabandi, A., Finn, C., & Levine, S. (2018). Deep Online Learning via Meta-Learning: Continual Adaptation for Model-Based RL. *arXiv:1812.07671*.

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

Regier, T., Carstensen, A., & Kemp, C. (2016). Languages support efficient communication about the environment: Words for snow revisited. *PloS one*, *11*(4), e0151138.

Regier, T., Kemp, C., & Kay, P. (2015). Word meanings across languages support efficient communication. *The handbook of language emergence*, 237–263.

Searle, J. R. (1969). *Speech acts: An essay in the philosophy of language*. Cambridge university press.

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

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

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

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

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

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

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