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 every new context. If we cannot rely on the existence of a shared dictionary for coordination, then, 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 of this work 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. Agents must integrate background expectations about reliable community-wide conventions (Lewis, 1969) with new *ad hoc*, partner-specific conventions constructed on the fly (Clark, 1996; Davidson, 1986). Chapters 1-2 formalize this proposal in a hierarchical Bayesian model and present simulation results capturing key effects from the literature. Chapter 3 evaluates an algorithm scaling this proposal to a recurrent neural network that adapts to human partners. Chapter 4 characterizes the dynamics of coordination in a large behavioral corpus of the classic Tangrams task. Chapter 5 tests how the communicative needs of the context shape the resulting *ad hoc* conventions. Chapter 6 assesses the generality of the proposed mechanisms by examining coordination in a *graphical* communication task.

, and developing computational cognitive models that give us mechanistic insight into how adaptive language use may be reverse-engineered from general principles of social cognition and learning. More broadly, this work opens avenues for future work to understand how social conventions are represented in the mind, how these conventions are shaped by our goals and environment, and how they give rise to the remarkable feats of social coordination that are so distinctive of human cultures (Hawkins, Goodman, & Goldstone, 2019).

Introduction

Do people adapt at all? Representations v. decisions, i.e. updating cognitively deep latent meanings v. priming/alignment.

Under this view, knowledge of community-wide linguistic conventions supplies a useful but uncertain first guess about the linguistic meanings that will be shared with a new partner. Through interaction, these general expectations are gradually tuned to the local context via social inferences based on that partner's language use. As each party learns about the other and attempts to be understood under their updated model of partner-specific meaning, they coordinate on a system of meaning that is tailored to be efficient and accurate for their present purposes. Finally,

Computational models of meaning must explain both the speaker's initial expectations about how words will be understood by novel partners and the dynamics of how these expectations may shift over the course of a particular conversation. In proposes a computational model that formalizes the problem of coordinating on meaning as hierarchical probabilistic inference, which I argue satisfies both of these conditions. Community- level expectations provide a stable prior, and dynamics within an interaction are driven by partner- specific learning.

Here we lay out the theoretical landscape and empirical evidence focusing on three key aspects of our theoretical approach: (1) the context-sensitive semantic priors that represent our initial uncertainty over what a novel partner is likely to mean by a word, (2) the path-dependence and increasing efficiency of our communicative behavior across even short periods of repeated interaction, and (3) the pattern of generalization from learning local interactions to new contexts and new partners.

This review also provides scaffolding for the experimental paradigm used in Chapters 2 and 4 and introduces the phenomena that we capture with our simulations in Chapter 3.

The core of this work is an inferential model of convention formation that explains flexible and adaptive language use across extended interaction as a consequence of hierarchical probabilistic learning. Through observing their partner's usage, agents attempt to infer and adopt their partner's underlying lexicon using global conventions as a prior. When both agents independently adopt such a learning strategy, they align to one another, coordinating on and implicitly creating new, shared conventions.

An inferential theory of convention formation

In **Chapter 2**, we propose a hierarchical Bayesian model of convention formation that formalizes our proposal (see Fig. 1): global conventions are learned and generalized over many extended interactions with many different people across a lifetime, and this shared semantic prototype is the backbone supporting rapid learning for new partners and situations.

What cognitive mechanisms explain these phenomena? Recent probabilistic modeling framework which derives pragmatic considerations of language use from more generic computational principles of social cognition in rational agents. For example, listeners treat the literal meaning of a question as a cue to infer and address the underlying goal of the

questioner **(Hawkins et al., 2015; ?, ?)**. And even though many labels may be literally true of an object, speakers refer to objects at an appropriate level of specificity given the other objects in context **(?, ?, ?)**. They also increase the specificity of their utterances under additional uncertainty over what else might be in their partners' visual field **(?, ?, ?)**. While powerful, these models are unable to explain the flexibility and adaptiveness of conventions over time, as they are constructed on top of *fixed* linguistic meanings.

A central aim of my dissertation work was to extend these models to capture how agents represent and rapidly coordinate on new conventions.I have implemented this theory in a Bayesian model that relaxes the assumption of fixed word meanings in probabilistic pragmatics models and instead treats them as dynamic and learnable for different partners. This captures a functional view of linguistic conventions as solutions to coordination problems, in which every agent is simultaneously seeking to infer the conventions that other agents are using, in order to support successful communication. This model successfully predicts coordination on idiosyncratic but path-dependent and stable form-meaning mappings in simulations, and shows that a preference for less costly utterances combined with learning gives rise to shorter descriptions over time **(Hawkins et al., 2017)**.

Here, we briefly present a sketch of the modeling approach. We begin by defining a lexicon as a function

$$\mathcal{L}_i : (w, o) \rightarrow [0, 1]$$

assigning any word-object pair a real-valued meaning in the unit interval. Just as our concept of a dog, built up over many individual experiences across a lifetime, provides stable expectations about the properties of a new instance – four legs, wagging tail, barking noises – our accumulated lexical knowledge provides stable communicative expectations. This knowledge is represented by a 'overhypothesis' or shared lexical representation Θ_0, which parameterizes the prior expectations about any individual partner's lexicon: $P(\mathcal{L}_i|\Theta_0)$.

Now that we have defined a hierarchical likelihood on lexical beliefs, we must say how we *learn* partner-specific models. Our beliefs about a particular partner's lexicon $\mathcal{L}_i$ are formed by integrating our abstract lexical knowledge Θ_0 with particular observations D_i of

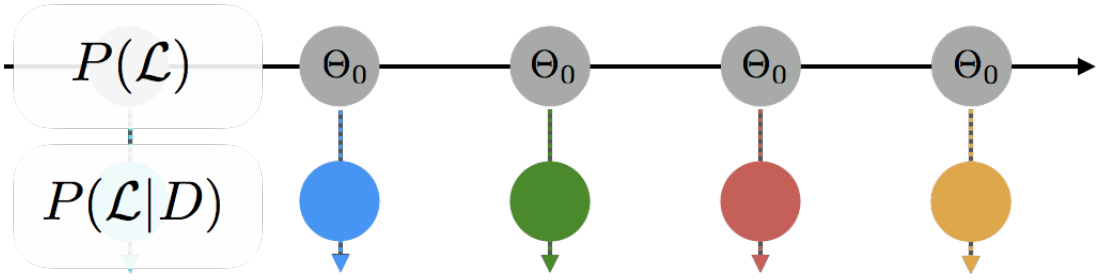

Figure 1: Schematic of our hierarchical model: Θ_0 parameterizes the agent's global beliefs about how their partner uses language, and then through interaction with each partner, they update a partner-specific lexical model based on data D from that interaction.

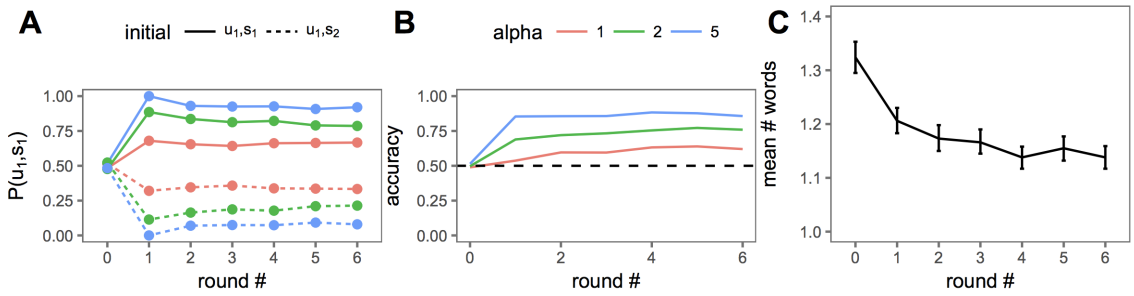

Figure 2: (A) Probability of speaker using one of two labels to refer to a state, broken out by initial observation: while players are initially ambivalent between the two labels (arbitrariness), the initial mapping is likely to persist (stability). (B) Accuracy rises as speaker and listener align. (C) When conjunctions are introduced into the grammar, utterances get shorter over time (reduction).

that particular individual:

$$P(\mathcal{L}_i|D_i) \propto \int_{\Theta_0} P(\mathcal{L}_i|D_i, \Theta_0)P(\Theta_0|D_i)$$

where the posteriors in the integral can be computed using Bayes rule:

$$P(\mathcal{L}_i|D_i, \Theta_0) \propto P(D_i|\mathcal{L}_i, \Theta_0)P(\mathcal{L}_i|\Theta_0)$$

While this holds when we only have observations from a single speaker, note that our posterior beliefs about Θ_0 are in fact informed by observations from *all* speakers: $D = \bigcup_{i=1}^{k} D_i$. Finally, to fully specify our model and compute our partner-specific lexical posterior $P(\mathcal{L}_i|D_i, \Theta_0)$, we must link our beliefs about a partner's lexica to their actual behavior with a likelihood function $P(D_i|\mathcal{L}_i, \Theta_0)$. This is naturally supplied by the Rational Speech Act framework: we assume speakers produce utterances that are parsimonious yet informative in context with respect to their lexicon, and listeners interpret utterances by inverting a speaker model. Because we expect our partner to use language rationally given some lexicon, the utterance they choose to refer to some object will be probable under some lexica and highly improbable under others. In this way, a particular agent's language use is a cue to their particular lexicon.

First, we show in simulations of a repeated interaction with a single partner that this model successfully allows coordination on arbitrary but path-dependent and stable form-meaning mappings and that a preference for less costly utterances combined with learning gives rise to reduction (see Fig. 6). Second, we test the generalization properties of our model by manipulating partners and contexts:

1. how 'sticky' are the pacts formed in one context (i.e. a Dalmatian in the context of other dogs) when the same target is transplanted to a new, less restrictive context (i.e. a Dalmatian in the context of a cat and bear; see Brennen & Clark, 1996),

2. to what extent do agents revert to longer utterances after swapping out partners midway through a game (see Wilkes-Gibbs & Clark, 1992; Metzing & Brennen, 2003; Yoon & Brown-Schmidt, 2014), and

3. how do local, partner-specific expectations generalize to global expectation over repeated interaction with multiple partners in the same community (see Fay et al. 2010).

1 Chapter 4: How do people coordinate on new conventions in naturalistic interactions?

A core function of language is *reference*: using words to convey the identity of an object in the environment. My research uses a rich, naturalistic communication paradigm called a *repeated reference game* that (1) requires participants to find some way of referring to novel, ambiguous stimuli that they don't already have strong conventions for and (2) asks them to refer to these same objects across multiple rounds as they build up a shared history of interaction, or 'common ground,' with their partner. I collected a large corpus ($> 15,000$ messages) of the natural language dialogue sent through a chat box by hundreds of pairs of participants playing this game over the web (Hawkins et al., 2017). I then used recent techniques developed from natural language processing to quantitatively characterize how the structure and content of participants' language evolves as they coordinate on new conventions.

First, going beyond classic work showing that messages get shorter and more efficient across repetitions, I extracted parts of speech and syntax trees from the natural text to understand *what* was reducing and *how*. I found that pairs systematically drop entire modifying phrases at each repetition, leaving only open-class parts of speech (e.g. an adjective and noun) by the final round. Second, I extracted word embeddings (e.g. GloVe vectors) for each message to calculate the similarity of messages within and across pairs. I found that 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. These findings provide higher resolution into the quantitative dynamics of convention formation: based on usage, existing words systematically acquire new meaning with a partner and support more efficient communication.

Lexical conventions are shaped by communicative context

These findings suggest that although there may be many equally effective solutions discovered by different pairs, the resulting conventions are nonetheless *adaptive*: they are shaped by the functional demands of communication in the local environment. **Chapter 5** tests this hypothesis using a controlled artificial-language paradigm adopted from studies of language evolution. As before, participants were paired to communicate about a set of target objects, but the statistics of the environment were manipulated across different pairs to make different distinctions relevant (Hawkins, Franke, Smith, & Goodman, 2018). In the 'fine' condition, distractors were very similar to the target, thus inducing a communicative need for finer distinctions. In the 'coarse' condition, distractors differed from the target at higher levels of a concept hierarchy. Using a post-test probing the resulting conventions and

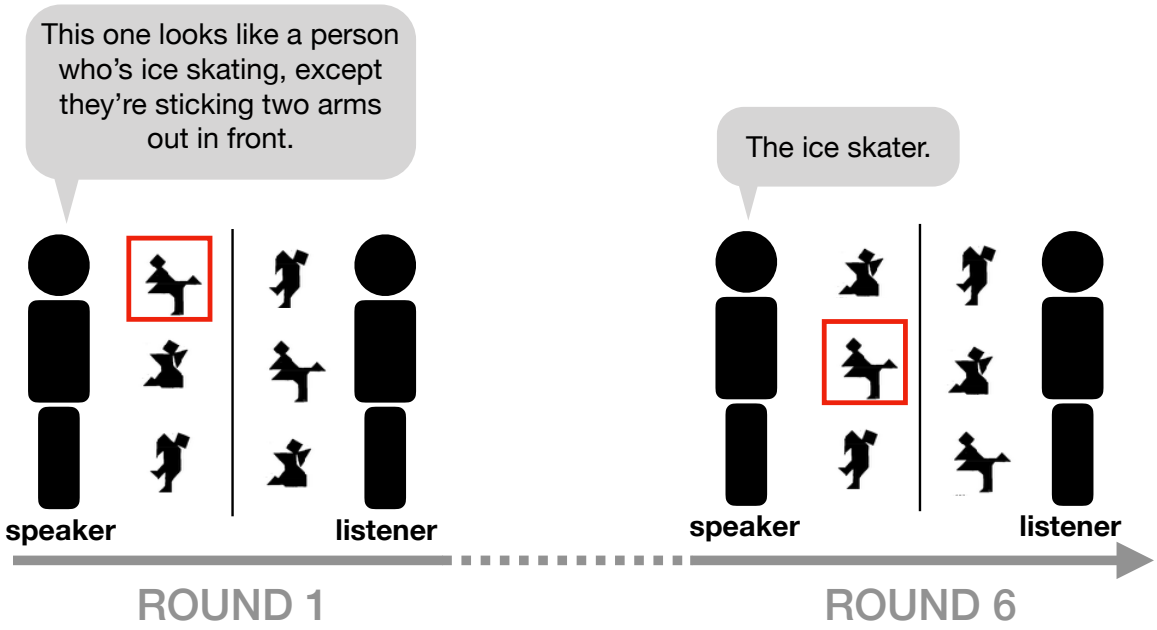

Figure 3: Generic setup for repeated reference game task in the lab using stimuli from Wilkes-Gibbs & Clark (1986); on every round, the speaker refers to each target in some context, and the listener attempts to pick out the intended referent. Both players are free to speak at any time.

a statistical approach inferring the lexicon at early stages from usage, I found systematic differences in the *level of abstraction* of conventions in these different contexts. These results indicate that pragmatic considerations of *informativity in context* shape the formation of lexical conventions over longer interactions.

Conclusions

Language is not some monolithic body of knowledge that we acquire at an early age and deploy mechanically for the rest of our lives. Nor is its evolution a slow, inter-generational drift. It is a means for communication – a shared interface between minds – and must therefore adapt over the rapid timescales required by communication. In other words, we are constantly learning language. Not just one language, but a large variety of related languages, across every repeated interaction with every partner.

My research program seeks to understand the core cognitive mechanisms that give rise to human social intelligence. In pursuit of this goal, I collect rich data from naturalistic social interactions in order to infer sets of candidate mechanisms, and instantiate them in computational cognitive models that reverse-engineer the algorithmic principles of social cognition. In addition to making quantitative predictions that address fundamental questions in cognitive science, these models also have broader application potential toward building more socially intelligent AI. By computationally characterizing these cognitive mechanisms, we will be better positioned to understand how interactions between individuals collectively

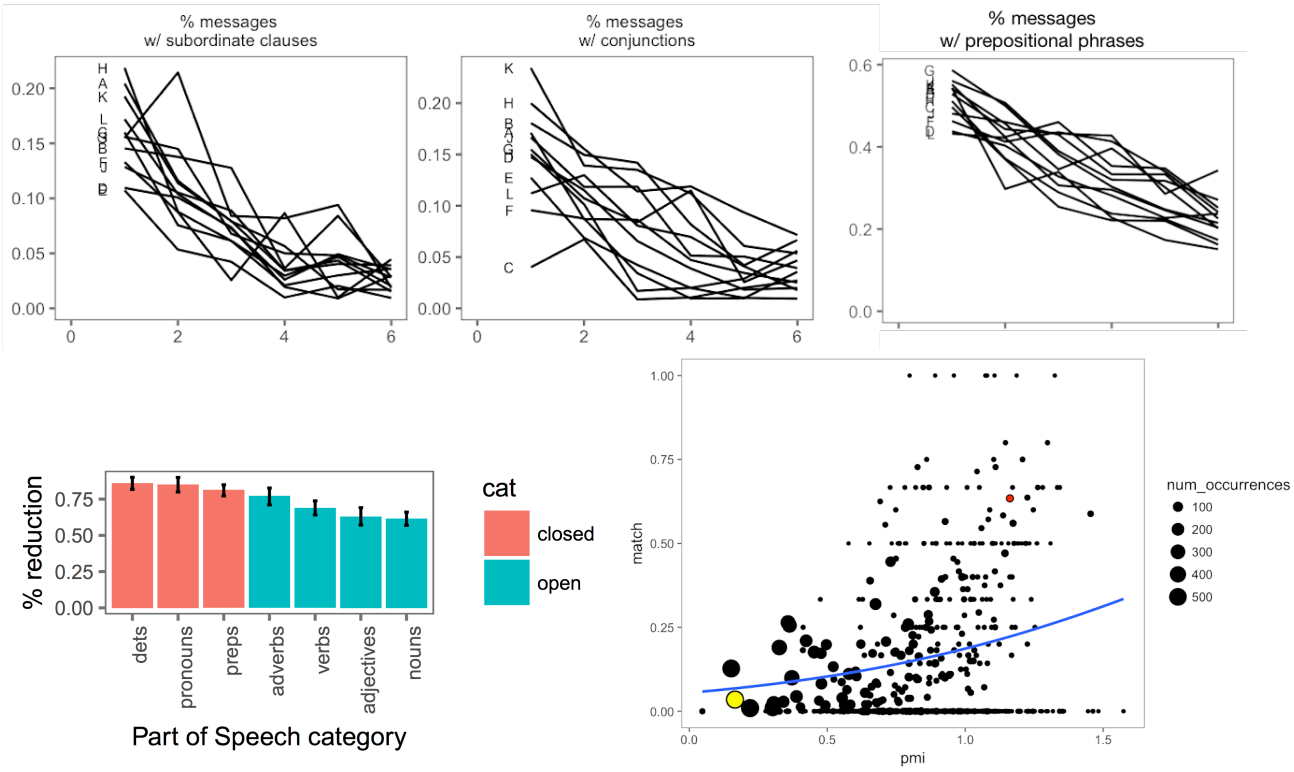

Figure 4: Reduction phenomena reproduced from Hawkins et al, 2017. Clockwise from left: (1) proportion of messages with subordinate clause, broken out by tangram, (2) proportion of messages with conjunction, (3) proportion of utterances containing prepositional phrase, (4) point-wise mutual information on first repetition predicts the likelihood of that word appearing in the final repetition (5) Reduction rates for different parts of speech. Error bars are bootstrapped 95% CIs

give rise to human culture.

References

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

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

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

Hawkins, R. 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., Stuhlmüller, A., Degen, J., & Goodman, N. D. (2015). Why do you ask? Good questions provoke informative answers. In *Proceedings of the 37th annual conference of the cognitive science society.*

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

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

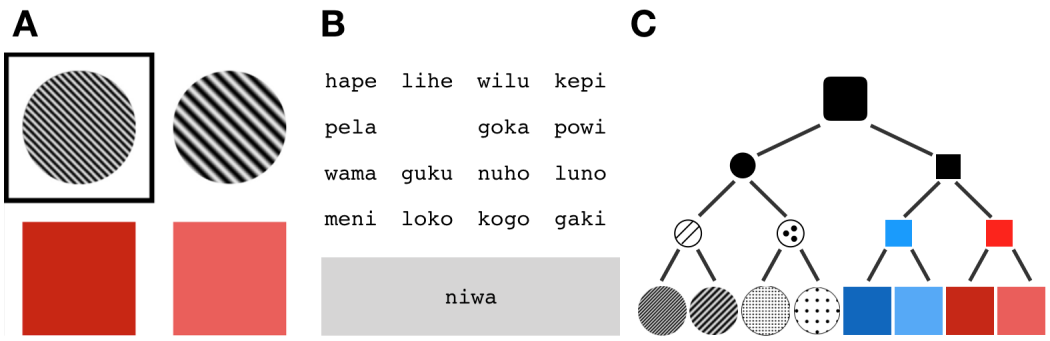

Figure 5: (A) Example of fine context where one of the distractors belongs to the same fine-grained branch of the hierarchy as the target (i.e. another striped circle), so any abstract label would be insufficient to disambiguate them. The target is highlighted for the speaker with a black square. (B) Drag-and-drop chat box interface. (C) Hierarchical organization of stimuli.

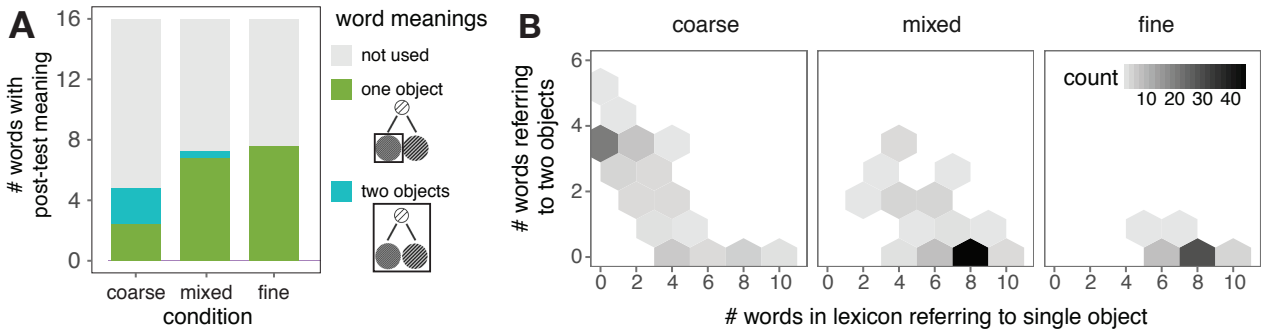

Figure 6: Pragmatic demands of context shape the formation of abstractions. (A) Mean number of words participants reported with specific meanings (applying to 1 object) or abstract meanings (applying to 2 objects). (B) Diversity of terms within reported lexica: many participants in the coarse condition reported a mixture of abstract and specific terms.