Coordinating on meaning in communication

A DISSERTATION PRESENTED
BY
ROBERT D. HAWKINS
TO
THE DEPARTMENT OF PSYCHOLOGY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN THE SUBJECT OF
PSYCHOLOGY

STANFORD UNIVERSITY
STANFORD, CA
MAY 2019

Advisor: Noah D. Goodman Robert D. Hawkins

Coordinating on meaning in communication

Abstract

How do we manage to understand each other, given that we are not telepathic? Human languages are a powerful solution to this challenging coordination problem. They provide stable, shared expectations about how the words we say correspond to the beliefs and intentions in our heads. However, to handle an ever-changing environment where we constantly face new things to talk about and new partners to talk with, linguistic knowledge must be flexible: we give old words new meaning on the fly. My dissertation investigates the cognitive mechanisms that support this balance between stability and flexibility. Chapter 1 introduces the overarching theoretical framework of communication as a meta-learning problem. Computational models of semantic 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. Chapter 2 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. Chapter 3 exploits recent connections between this hierarchical Bayesian framework and continual learning in deep neural networks to propose and evaluate a computationally efficient algorithm implementing this same model at scale in an adaptive neural image-captioning agent. In Chapter 4, I provide an empirical basis for further model development by quantitatively characterizing convention formation behavior in a new corpus of natural-language communication in the classic Tangrams task. By using techniques from natural language processing to examine the (syntactic) structure and (semantic) content of referring expressions, we find that pairs coordinate on equally efficient but increasingly idiosyncratic solutions to the problem of reference. Chapter 5 uses an artificial-language reference game paradigm to test the hypothesis that communicative context systematically shapes which conventions form. Finally, Chapter 6 investigates the generality of the proposed computational mechanisms by examining convention formation in a *graphical* communication task. Taken together, this line of work builds a computational foundation for a dynamic view of meaning in communication.

Contents

1 Introduction 1
- 1.1 Using repeated reference games to study convention formation in the lab 4
- 1.2 Mechanism #1: A probabilistic lexicon 6
- 1.3 Mechanism #2: Rapid lexical learning 10
- 1.4 Mechanism #3: Generalization 16
- 1.5 Discussion 25

2 An inferential model of convention-formation 27
- 2.1 Adapting to a single partner 29
- 2.2 Generalizing to new partners 33

3 Continuous adaptation for efficient machine communication 40
- 3.1 Approach 42
- 3.2 Evaluations 47
- 3.3 Analysis 49

4 Characterizing conventions: the dynamics of structure and content 51
- 4.1 Methods: Repeated reference experiment 53
- 4.2 Results: characterizing the dynamics of content 58
- 4.3 Results: characterizing the dynamics of structure 69
- 4.4 Discussion 75

5 Emerging abstractions: Lexical conventions are shaped by communicative context 78
- 5.1 Experiment: Repeated reference game 81
- 5.2 Model-based Analysis 86
- 5.3 Discussion 90

6 Disentangling contributions of visual information and interaction history in the formation of graphical conventions 93
- 6.1 How does repeated reference support successful visual communication? 95
- 6.2 What explains gains in efficiency? 100
- 6.3 How do visual features of drawings change over the course of an interaction? . . . 104
- 6.4 Discussion 107

7 Conclusion 110

References 139

To my mother and father. My education would not have been possible without your sacrifices, and I aspire to live with your humility, generosity, and curiosity.

Acknowledgments

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Morbi commodo, ipsum sed pharetra gravida, orci magna rhoncus neque, id pulvinar odio lorem non turpis. Nullam sit amet enim. Suspendisse id velit vitae ligula volutpat condimentum. Aliquam erat volutpat. Sed quis velit. Nulla facilisi. Nulla libero. Vivamus pharetra posuere sapien. Nam consectetuer. Sed aliquam, nunc eget euismod ullamcorper, lectus nunc ullamcorper orci, fermentum bibendum enim nibh eget ipsum. Donec porttitor ligula eu dolor. Maecenas vitae nulla consequat libero cursus venenatis. Nam magna enim, accumsan eu, blandit sed, blandit a, eros.

1
Introduction

We do not have to learn language from scratch with every new person we meet. A native of Chicago can try out a coffee shop in San Francisco without needing to laboriously work out with the barista a brand-new way of ordering an 'espresso,' and a 21st century reader can largely make sense of a 19th century novel without any personal contact with its author. This degree of stability across geography and time makes language indispensable for coordination in a social species: everyone who belongs to a language community assumes that others will share at least some common beliefs, or *global conventions*, about what words mean (Lewis, 1969). In this sense, the ability to competently generalize to novel communicative partners in novel contexts is what it means to be fluent in a language.

At the same time, no two speakers of a language share exactly the same lexicon, and to make matters worse, speakers seem to constantly come up with new expressions and senses on the fly (Davidson, 1986; H. H. Clark, 1998). Drop into any conversation between friends and you'll be wading in a stream of shorthand, lingo, slang, references, and inside jokes — some of which you might understand, but the rest of which may be meaningful to them alone. While it is tempting to think of word meanings as residing only in dictionaries, we instead find ourselves continually negotiating new meanings for old words to communicate complex thoughts, intentions, and beliefs in context. Across repeated interactions with a particular partner or group or partners, we build up intricate, idiosyncratic models: not just about how *we* as English speakers collectively use language, but how *this person* is expected to use language. We learn the contours of their speech and form *local conventions* from our shared history. Complex stories or ideas that took long discursive conversations to initially cover can be referred back to using a brief turn of phrase. Most strikingly, this adaptation can take place to different degrees over any time-scale: from years of close scientific collaboration to a few minutes in a doctor's office.

A core puzzle for cognitive science, then, is reconciling the overarching stability of our linguistic representations with our remarkable flexibility in coordinating on new meanings. What do we do when our global conventions aren't sufficient — when we have to talk about something we've never had to talk about before with a partner we've never met? How do we adapt so quickly? And how do we determine which learned meanings we can expect to stably generalize to new contexts or partners, and which only hold in the narrow scope of one partner?

To address this puzzle, we begin by breaking down the computational challenge of coordinating on meaning into three distinct cognitive mechanisms. While this dissertation will primarily focus on the second of these mechanisms, we will argue that all three are tightly intertwined in supporting the stability and flexibility of communication.

1. Prior expectations: When we first encounter a new communication partner in a new context,

we call upon some representation about what we think different signals mean to them. This representation of meaning must be sensitive to the overall statistics of the population: more people are familiar with the use of *dog* to refer to the beloved pet than *sclerotic aorta* to refer to the potentially dangerous health condition. It must also be sensitive to the immediate context of the interaction: a cardiologist should have different expectations about a novel colleague than a novel patient.

2. Rapid adaptation: Within a few minutes of conversation, we can considerably strengthen our expectations about our partner's lexicon based on earlier utterances and feedback, and adjust our own usage accordingly. For example, even if we are not initially familiar with the term *sclerotic aorta*, a few minutes spent discussing the condition in simpler terms should make us more confident using the term with that partner in the future. This social learning mechanism must allow for signal *reduction* – simpler, more efficient ways of referring to the same thing over time – and *path-dependence*: early reinforcement of certain meanings increases their later usage, however arbitrary or provisional they began.

3. Generalization: When we encounter the same partner in a new context, we should expect some 'stickiness' from previous learning. Language does not reset at context boundaries. In addition, the lexical model we've learned within a conversation should be largely *partner-specific*. Just because we now expect Partner A to be familiar with a *sclerotic aorta* shouldn't radically change our expectations about Partner B. Over enough interactions with different language users, however, our initial representations should be able to shift to take these data into account. To generalize appropriately, we must be able to correctly attribute whether a usage is idiosyncratic to a particular speaker, or a global convention we should expect to hold across the whole community.

In the remainder of Chapter 1, we consider the empirical evidence supporting the role each of these three core competencies, reinterpret this evidence from a computational perspective, and discuss several broader implications. Though more wide-ranging sources of data could potentially be relevant, we restrict our present scope to a family of interactive communication experiments called *repeated reference games* that will be used extensively in subsequent chapters. This task provides a natural and productive paradigm for studying how people coordinate on meaning in the lab.

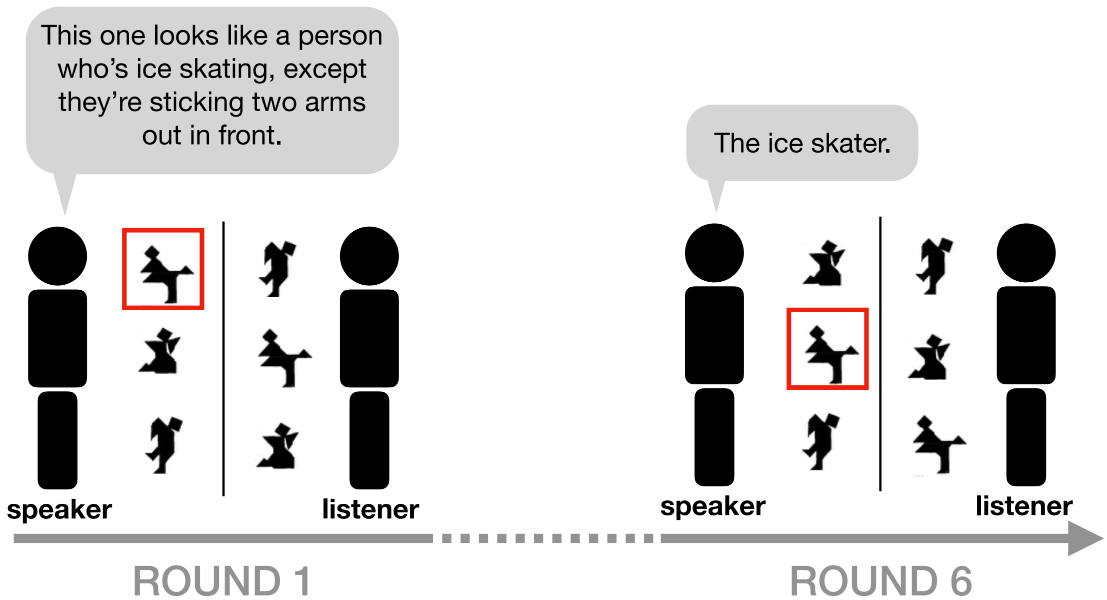

Figure 1.1: Generic setup for repeated reference game task in the lab using stimuli from Wilkes-Gibbs & Clark (1986); on every round, the speaker refers to each target in some context, and the listener attempts to pick out the intended referent. Both players are free to speak at any time.

1.1 Using repeated reference games to study convention formation in the lab

In their simplest design (see Fig. 1.1), pairs of participants are shown arrays of objects, presented in randomized order. On each round of the game, one player – the speaker – must produce a message allowing their partner to select a given target from the context. By fixing a closed set of referents and a clear communicative goal in common ground, these games vastly simplify the tangle of real-world communicative behavior. At the same time, by allowing real social partners to freely interact using natural language, we avoid the hazards of artificial or confederate-based language tasks (Kuhlen & Brennan, 2013) and expose the rich dynamics of language use in context. Given the current state of the field, then, reference games are arguably an ideal compromise between analytic tractability and ecological validity.

While one-shot reference games, which manipulate context, have been instrumental in revealing

	Parameter	Example parameter settings
Partner design	What feedback is provided?	- no feedback at all - only correct/incorrect - real-time responses from partner
	Are you playing with the same partner?	- same partner for whole game - swap out partners every round - swap after k rounds
	What do you know about your partner?	- anonymous stranger - stranger with perceptual information - close friend
	How consistent are roles across repetitions?	- consistent director/matcher - alternate roles each round
Stimulus design	How familiar are targets?	- very familiar: colors, household objects - not at all familiar: tangrams, novel line drawings
	How complex are targets?	- very complex: busy visual scenes, clips of music - not at all complex: geometric drawings
	How consistent are targets across repetitions?	- exact same image of object - different pose/view of same object - different objects from same neighborhood
Context design	How similar are distractors to the target?	- very similar: same basic-level category - not at all similar: other categories
	What is the size of context?	- between 2 and 21
	How consistent is context across repetitions?	- exact same context each round - randomized context (sometimes far, sometimes close)
Repetition design	How many repetitions per target?	- between 3 and 100
	What is spacing between repetitions?	- block structure - sequential structure with interspersed contexts
Modality design	What medium is used for communication?	- text - audio - gesture - drawing

Table 1.1: Proposed parameterization for repeated reference games, each of which theoretically impacts the formation of conventions.

systematic pragmatic effects in the referring expressions people tend to generate (Krauss & Weinheimer, 1967; Koolen, Gatt, Goudbeek, & Krahmer, 2011; Graf, Degen, Hawkins, & Goodman, 2016; van Deemter, 2016), a range of even richer phenomena begin to emerge when the communication task is *repeated* and the same target must be referred to multiple times. Since Krauss and Weinheimer (1964) first attempted such a design, many variations on this basic setup have been designed to test the boundaries of adaptation, manipulating the kinds of objects used as targets, the contexts in which the objects appear, the identity of one's partner across repetitions, the feedback available, and the medium participants use to communicate. In Table 1.1, we propose a potential parameterization of this family of repeated reference games, suggesting a set of conditions controlling when conventions may form. We organize our review of this literature along the three core challenges observed earlier, which roughly correspond to the temporal structure of a repeated reference game. In experimental terms: What influences the content of initial messages, how do these messages change over the coarse of the game, and under what conditions do these changes transfer to other scenarios?

1.2 Mechanism #1: A probabilistic lexicon

People know a lot of words (Bergelson & Aslin, 2017). Exactly how flexible knowledge about words and their meanings is structured and accessed in one's own "mental lexicon" remains a significant open question in cognitive science (Jones, Willits, Dennis, & Jones, 2015; Griffiths, Steyvers, & Tenenbaum, 2007; Huth, de Heer, Griffiths, Theunissen, & Gallant, 2016; N. D. Goodman & Lassiter, 2014). For the purpose of communicating with another speaker, however, a more relevant question is what lexicon we think our *partner* is using. In this section, we review evidence from the initial rounds of repeated reference games that these lexical expectations are *probabilistic* and *context-sensitive*, thus providing a basis for interpreting expectations about a novel partner's lexicon as a probabilistic prior $P(\mathcal{L}_i|\Theta_0)$.

1.2.1 Uncertainty in lexical expectations

While it is convenient to view the lexicon as fixed knowledge (Cruse, 1986; Pinker, 1995; Frank & Goodman, 2012), meanings are in reality quite flexible and ad hoc (H. H. Clark, 1983; H. H. Clark & Gerrig, 1983; E. V. Clark & Clark, 1979; Gerrig & Bortfeld, 1999; Lascarides & Copestake, 1998;

Glucksberg & McGlone, 2001; Lassiter & Goodman, 2015). Correspondingly, recent computational models have explored the possibility that we instead represent semantic *uncertainty* over which meanings our partner might intend (e.g. Cooper, Dobnik, Larsson, & Lappin, 2015; Bergen, Levy, & Goodman, 2016; Smith, Goodman, & Frank, 2013; Potts, Lassiter, Levy, & Frank, 2016; Hawkins, Frank, & Goodman, 2017). This uncertainty leaves its signature on the initial round of repeated references games, when speakers are attempting to produce descriptions of potentially ambiguous objects for a novel partner.

In one direct demonstration, Fussell and Krauss (1989a) asked forty students to produce referring expressions for abstract line drawings in a repeated reference game set-up. Instead of proceeding to play the game in person, however, participants were told that their messages were intended for later identification (see Krauss, Vivekananthan, & Weinheimer, 1968; Danks, 1970; Innes, 1976, for earlier variations on this design). Half the participants were told that these messages were intended for *themselves* in the future (the 'non-social' or 'familiar listener' condition) while the other half were told that an anonymous other would see them (the 'social' or 'unfamiliar listener' condition).

Because both groups were faced with the same communicative task, this manipulation provided some evidence about how lexical expectations differ depending on the listener. If participants used the same fixed meaning when reasoning about themselves and others, we would expect similar messages. Instead, utterances intended for others were more than twice as long as utterances for oneself (12.7 vs. 5.0 words). Furthermore, these social expectations supported effective communication: when participants were brought back into the lab 3-6 weeks later to perform a 30-way identification task given these previously collected descriptions, they performed best given their own (86% accuracy) but when presented with others participants' descriptions, they did significantly better when those descriptions were explicitly designed for an unfamiliar listener (60% vs. 49%).

Why produce longer utterances for others? A key empirical observation is that similar self-other length effects were found by Innes (1976) using ambiguous stimuli like abstract designs, inkblots, and poems, but *not* by Krauss et al. (1968) where the same procedure was conducted with familiar color chips; Hupet, Seron, and Chantraine (1991) make this dimension of the stimuli explicit by independently norming the 'codability' of a large array of tangrams and showing longer other-directed messages for more ambiguous stimuli (though they didn't run a 'self' condition).

One parsimonious explanation is that in contexts where global conventions are stronger and lexical uncertainty is lower (e.g. for common colors), speakers expect others to share identical lexical beliefs and can get away with similarly terse descriptions for self and other. Meanwhile, for ambiguous stimuli like inkblots or tangrams that speakers have had limited experience communicating about, they have substantial uncertainty about how an anonymous other, drawn from their global prior, will interpret their words. It could therefore be worth spending a few additional words to provide clarifying information, saying "the upside-down martini glass in a wire stand" instead of just "the martini" (Hawkins et al., 2017).

This explanation is also consistent with broader linguistic phenomena outside the realm of repeated reference games. For example, Potts and Levy (2015) showed that lexical uncertainty is critical for capturing constructions like *oenophile or wine lover*, where a disjunction of synonymous terms is taken to convey a definition – information about the lexicon – rather than a disjoint set. While the reasons that speakers produce such constructions are surely more complex, we suggest the further conjecture that speakers are more likely to produce this definitional *or* when the component word is rarer or more obscure: when there is additional uncertainty over its likely meaning in the listener's lexicon.

1.2.2 Context-sensitivity in lexical expectations

A global lexical prior $P(\mathcal{L}_i|\Theta)$ is an excellent guide to what meanings an anonymous member of our language might have. Yet we typically know a thing or two about a novel conversation partner before we start talking to them — their perceived age, gender, dress, social role, any behavior we've witnessed (Davidson, 1986; Kleinschmidt & Jaeger, 2015). Our prior should be flexible enough to take this evidence into account. Repeated reference games in the lab usually take care to disguise as much of this evidence as possible, though some properties of the sample are unavoidable: when recruited on a college campus, participants can safely assume their partner is another student, for instance.

In one study, Fussell and Krauss (1992) tested the extent to which lexical priors are well-calibrated to minimal expectations of cultural knowledge in the target population. In a repeated reference game using faces of public figures like Woody Allen and Ronald Reagan, they found that speakers gave lengthier initial descriptions for figures who were expected to be less identifiable or well-known, as estimated from independently elicited priors. This relationship held when restricted to messages containing correct names, meaning that speakers who themselves knew the identity of the figure were nonetheless more likely to add additional information to their initial description when they expected a typical partner not to know.

Fussell and Krauss (1992) also predicted that additional information about the *gender* of partners would be incorporated into lexical expectations. In particular, they expected that because men and women stereotypically have some gendered lexical expertise (e.g. men know more names for car parts), participants would be more likely to provide additional information to the opposite gender. They again found an effect of overall expected familiarity, but did not find an interaction with gender, musing that the perceived discrepancy was perhaps too small to detect in the items used.

It is natural to view the self and the stranger as points on a continuum, with stronger initial expectations for close friends or family with whom we have a long history of interaction and potentially weaker expectations for children, non-native speakers, or out-group members. This first prediction was tested by Fussell and Krauss (1989b) who brought self-identified pairs of friends into the lab and had them individually produce descriptions such that 'their friend' could identify it. They failed to find a significant difference in description length from the descriptions produced for strangers in

Fussell and Krauss (1989a), but this negative result is somewhat hard to interpret for two main reasons acknowledged by the authors. First, their interpretation of 'friendship' was not well-controlled and many pairs were only casual acquaintances drawn from the same college population as the 'other student' that participants in the 'stranger' condition were instructed to produce descriptions for. Second, even with deep knowledge of an intimate partner's lexicon, it is not clear how relevant this knowledge would describing for a set of abstract line figures: it was specifically designed to be novel and somewhat unnatural.

Despite these underwhelming early results, the context-sensitivity of lexical priors remains a tantalizing area to revisit. Sources of variance in lexical expectations across speakers (King & Sumner, 2015), and mechanisms supporting speaker-specific expectations (Tesink et al., 2009; Van Berkum, Van den Brink, Tesink, Kos, & Hagoort, 2008) may be more amenable to study using modern tools. In particular, more recently developed methods for measuring subjective beliefs and expectations (e.g. Franke et al., 2016; Delaney-Busch, Morgan, Lau, & Kuperberg, 2017) could provide much more direct access to underlying lexical beliefs, and large-scale online experiments could more systematically uncover the underlying structure of social group representations along which lexical expectations are organized.

1.3 Mechanism #2: Rapid lexical learning

If our lexical priors – our global conventions – serve as a source of stability in meaning over longer timescales, then what accounts for our extraordinary flexibility over short timescales? How do we coordinate on efficient local conventions, or *conceptual pacts*, for talking about things we've never talked about before? In this section, we review the dynamics of coordination within repeated reference games and explore the possibility, formalized in Chapter 2, that rapid adaptation can be understood in a Bayesian modeling framework as lexical inference given partner-specific data.

1.3.1 Convergence on efficient conventions

The most well-known phenomenon in repeated reference games is a reduction in message length over multiple rounds. Krauss and Weinheimer (1964) were the first to report this phenomenon in a short technical report introducing the repeated reference game paradigm, and it has been replicated many times under many conditions (most notably by H. H. Clark & Wilkes-Gibbs, 1986, in a much more streamlined experimental design using tangram shapes). Out of historical interest, it is worth describing the original design in detail.

Both players were given an identical set of 6 cards marked randomly with 1-6 and A-F, respectively, containing the same 6 drawings in different orders. The pair's goal was to figure out the correspondences between their 6 cards by talking about the locations of the images. In each set, three images were 'redundant,' appearing in the same location on every card—discussing these was not very useful for the task—while the other three were 'diagnostic' and necessarily had to be referred to. This design therefore had the peculiar property that different drawings appear with different frequencies: some objects were referred to nearly 100 times (e.g. if diagnostic for every set of cards across all 16 rounds) and others only a handful of times.

Their core descriptive result was that, taken in aggregate, frequently mentioned targets tend to be labeled using shorter phrases than infrequently mentioned targets, thus reproducing Zipf's law within the microcosm of a single conversation. To explain the process by which such a distribution emerges, they reasoned that labels may change with repeated use over the course of interaction. Indeed, the first time participants referred to a figure, they used a lengthy, detailed description ("the upside-down martini glass in a wire stand") but with a small number of repetitions – between 3 to 6 times, depending on the pair – the description was reduced down to the limit of just one or two words ("martini").

Note that although initial messages are just as long or longer than the other-intended messages collected by Fussell and Krauss (1989a), final messages are as short or shorter than the one-shot messages intended for *oneself*. Furthermore, final messages are often incomprehensible to overhearers who were not present for the initial messages (Schober & Clark, 1989). This observation sets up the central empirical puzzle of convention formation: how does a short word or phrase that would have been completely ineffective for communicating under the initial lexical prior become perfectly understandable over mere minutes of interaction? What changes inside participants' minds in the interim?

One simple non-social explanation — that reduction is merely an effect of familiarity or repetition on the part of the speaker — can be easily dispelled. When participants are asked to repeatedly refer to the same targets for a hypothetical partner, no reduction is found, and in some cases utterances actually get longer (Hupet & Chantraine, 1992). Whatever is changing must be a result of the *interaction* between partners. An alternative explanation suggested by our probabilistic model is that reduction is driven by lexical learning as communication partners coordinate on ad hoc names. If long initial messages can be explained as the result of initial uncertainty in the lexical prior, as discussed in the previous section, then a decrease in uncertainty licenses shorter messages (Hawkins et al., 2017).

1.3.2 Signatures of reduction

What are empirical cues to this reduction in uncertainty? The first is the use of *hedges*. Hedges are expressions like *sort of* or *like*, and morphemes like *-ish*, that explicitly mark uncertainty or provisionality, such as *a car, sort of silvery purple colored* (Brennan & Clark, 1996; Fraser, 2010; Medlock & Briscoe, 2007). If participants reduce their lexical uncertainty over successive rounds, then we might expect a corresponding decrease in explicit markers of this uncertainty. Brennan and Clark (1996) counted hedges over four repetitions of an initially ambiguous target and found widespread use of *hedges* on the first round (occurring 26% of messages) but almost complete absence on the last (only 2% of messages). They also found very few initial hedges for targets with low initial uncertainty (e.g. a shoe in the context of dogs and fish), providing additional evidence for the role of *lexical* uncertainty as opposed to a generic social use of hedges.

Another characteristic of uncertainty reduction lies in *what* gets reduced, which we discuss in depth in Chapter 4. Is the speaker adopting a fragment shorthand by randomly dropping function words, or are they simplifying or narrowing their descriptions to names by omitting redundant details? Closed-class parts of speech like determiners and prepositions *are* much more likely to be dropped than open-class parts of speech like adjectives and nouns. But when we examine broader grammatical units using recent NLP techniques, we find that entire modifying clauses are increasingly likely to be dropped (Hawkins et al., 2017). This accords with early hand-tagged analyses by Carroll (1980), which found that in three-quarters of transcripts from Krauss and Weinheimer (1964) the short names that participants converged upon were prominent in some syntactic construction at the beginning, often as a head noun that was initially modified or qualified by other information.

These more fine-grained analyses suggest that reduction is grounded in the prior lexical content of the interaction and the speaker's increasing confidence in how the listener will interpret an initially ambiguous label. Like the evidence we reviewed about lexical priors, however, this evidence remains indirect and raises the need for more careful, direct measurement of lexical uncertainty over interaction.

1.3.3 Quality of feedback

If adaptation is learning, then the extent to which partners adapt should depend critically on the quality of the data D_i on which they are conditioning: $P(\mathcal{L}_i|\Theta, D_i)$. In the absence of additional cues to the meanings that their partner is using to interpret their messages, a speaker or drawer can only continue to rely on their prior, or indeed elaborate upon it. A common feature of the reference games reviewed so far is the capacity for *real-time feedback channel*: either player may say any-thing at any point in time, thus allowing for interruptions, back-channel responses (uh-huh, hmmm, huh?), clarification questions, and so on. To what extent is this design choice necessary for reduction? Krauss and Weinheimer (1966) were the earliest to address this question by manipulating the kind of feedback received by the speaker.

Intuitively, we might expect that if the speaker is unsure how their longer descriptions are being interpreted – unsure whether or not they can get away with shorter, more ambiguous expressions – they may not have enough evidence about meanings to justify shorter utterances. Indeed, Krauss and Weinheimer (1966) found that even when told that their partner was getting 100% correct, entirely blocking the verbal feedback channel significantly limited the reduction effect. Speakers converged to utterances that were about twice as long – twice as inefficient – in the limit. Telling speakers that their partner was performing poorly also inhibited reduction as a main effect, though to a lesser extent. In the extreme case of trying to communicate to a listener who can't respond and appears to not understand, speaker utterance length actually increased with repetition after an early dip. Hupet and Chantraine (1992) later found that in the *complete* absence of feedback — when the speaker is instructed to repeatedly refer to a set of objects for a listener who is not present and will do their half of the task offline — there is also no reduction in message length. On the listener's part, too, the ability to actively *give* feedback appears critical for learning. Schober and Clark (1989) showed that listeners who overheard the entire game were significantly less accurate than listeners who could directly interact with the speaker, even though they heard the exact same utterances.

More graded disruptions of feedback seem to force the speaker to use more words overall but not to significantly change the rate of reduction (though rigorous comparisons between rates have not been conducted). For example, Krauss and Bricker (1967) tested a transmission delay to temporally shift feedback and an access delay to block the onset of listener feedback until the speaker is finished. Later, Krauss, Garlock, Bricker, and McMahon (1977) replicated the adverse effect of delay but showed that undelayed visual access to one's partner cancelled out the effect and returned the number of words used to baseline.

1.3.4 Why so fast?

Some access to minimal feedback from one's partner therefore appears to be a necessary condition for convention formation. Without it, there is no reliable cue to the partner's lexicon; no lexical learning can take place, and consequently no social coordination. Yet this condition alone doesn't explain the *speed* with which partners adapt, approaching 'one-shot' learning. Three additional factors seem relevant.

First, like other scenarios of rapid learning from sparse data, abstract prior knowledge is crucial (Tenenbaum, Kemp, Griffiths, & Goodman, 2011; Lake, Ullman, Tenenbaum, & Gershman, 2017): agents do not start from scratch, they must only fine-tune their pre-existing global conventions to fit their immediate partner and context. Second, agents have pragmatics on their side. In the RSA model linking lexical knowledge to behavior, listeners assume their partner is attempting to be *informative* and pragmatic speakers, in turn, *expect* listeners to do so. These assumptions dramatically strengthen feedback. Because listeners reason about alternatives – that the speaker *would* have used another word or description if it described the target better in context – both agents actually learn about the meanings of words that were not uttered*. Similarly, the existence of a listener backchannel (knowing a listener *would* object or ask for clarification if their lexicon differed) implies evidence for the utterance's meaning in the absence of such objections. A third factor is the sociolinguistic information derived from social group inferences, which are often tightly controlled in lab settings but likely more relevant in the real world.

* This is the *lateral-inhibition* dynamic described at length by Steels (2003); Steels and Belpaeme (2005); Steels (2015), which emerges naturally in our model from basic Gricean principles.

1.4 Mechanism #3: Generalization

If the local conventions – or pacts – formed over the course of an interaction reflect learning, then what are the boundaries of that learning? How does meaning in one context with one partner transfer to expectations about new contexts and new partners? In this section, we discuss three empirical properties of generalization that computational models of convention formation must account for. Over short time scales, lexical pacts are *partner-specific* but stable across *contexts*. Over longer time scales, however, as agents repeatedly interact with multiple partners in larger social networks, pairwise conventions generalize to global conventions expected to be shared across the entire *community*. A mechanism for generalization, then, is not only key to understanding the appropriate scope of local conventions, but also where our global conventions came from in the first place.

1.4.1 Partner-specificity

One of the most salient and well-studied properties of local conventions is precisely that they *don't* generalize immediately to novel partners. In other words, the speaker is learning a specific model of their partner on the basis of shared history, not just privately forming an association between words and objects. This can be demonstrated in a repeated reference game by swapping in a novel parter after several rounds of interaction (Wilkes-Gibbs & Clark, 1992; see also Brennan & Clark, 1996, Weber & Camerer, 2003, Yoon & Brown-Schmidt, 2014 for variations on this design). If the speaker's lexical representation did not distinguish between partners with different histories, then we should expect no difference before and after this intervention. Instead, Wilkes-Gibbs and Clark (1992) found that speakers immediately reverted to longer messages—a 275% increase—bringing them nearly back to the length of the initial messages used with the original partner.

Other interventions explore intermediate cases, where the swapped-in listener was not *entirely* novel. For example, when Wilkes-Gibbs and Clark (1992) introduced them as a "silent participant" during the first phase, sitting at the same table as the speaker and able to make eye contact, there was only a 67% increase in message length. When they were an *omniscient bystander*, observing the audio and visuals of the entire exchange from a separate room, length increased slightly more (100%). Finally, when they were a *simple bystander*, seated some distance behind the director such that they could not be monitored or see the tangrams being referred to, length increased as much as with a fully novel partner.

These graded effects reflect degrees of *partial information* about the swapped-in listener's beliefs, posing a challenge even for computational models that allow partner-specific learning. It is simple for a Bayesian learning account to explain why speakers had no particular expectations about the *simple bystander*, who could not see the tangrams that were being referred to and therefore was not provided with the relevant data to learn meanings. But for third parties who *did* have full access to the interaction, whatever additional information the speaker is using cannot be limited to the assumption that the swapped-in listener can consult the same data, otherwise the *omniscient bystander* should be treated the same as the *silent participant*.

One explanation, which could plausibly be incorporated into a model, is that although the silent participant cannot verbally respond, they do nonetheless provide a minimal feedback channel through their physical presence, e.g. eye contact or body language indicating that they are paying attention. Because they are co-present at the table, they belong to the intended audience for the speaker's messages, and can be monitored for cues to understanding; this evidence in turn may not indicate partner generalization per se, but joint learning about the two listeners.

A related question concerns how partner identity is cued or inferred: even for a model that learns partner-specific lexical representations, the identity of a particular partner may be noisy, requiring inference under uncertainty about what is in common ground. In a challenging task used by Horton and Gerrig (2002), a speaker played several rounds of a repeated reference game with two listeners at once. Critically, each listener only had access to a *subset* of the total grid of targets in front of the speaker, which the speaker could only learn through interaction. Having converged on local conventions for one set of objects with listener A and for another set of objects with listener B, the speaker was then put in a test phase where they had to communicate *all* objects with each listener separately. As a main effect, Horton and Gerrig (2002) found a small boost in message length for the subset of objects that *weren't* originally part of the present listener's array, indicating that the speaker flexibly shifted their messages to accommodate the differing history with each listener. There was also evidence from intra-utterance analyses that partner-specific representations were used to form the very first messages and the effects weren't simply driven by low-level feedback (e.g. if the speaker began by producing the reduced label used with the previous partner, and only added additional information after the listener indicated trouble understanding).

This result by itself is a small but compelling addition to the cumulative evidence of partner-specificity, but also presents two additional challenges for computational models related to the noisiness of partner-specific learning. First, there was a strong order effect, with a much stronger increase in message length at the *second* test session (with partner B), suggesting some additional learning of what was in common ground with partner B in the first test session with partner A. Second, in a follow-up study, Horton and Gerrig (2005) found that clustering the stimuli to facilitate easier learning of which objects were known by which speakers during the first phase significantly strengthened the partner-specific effect. These observations indicate that under challenging conditions, learning partner-specific representations is a noisy affair and often appears far from 'optimal' even in a rational learning model.

Finally, while most repeated reference games have focused on testing speaker adaptation, there is evidence that partner-specific information is represented by the listener, too. For example, Metzing and Brennan (2003) tested the partner-specificity of listener representations in an eye-tracking study with a confederate *speaker*. In one condition, after several rounds of interaction "converging" on a pre-scripted convention (e.g. *"the shiny cylinder"*), the speaker *broke* the pact, suddenly introducing a novel but synonymous term (e.g. *"the silver pipe"*). In another condition, the speaker was replaced such that the novel term was produced by a novel speaker. In this latter case, the listener looked at the target just as quickly as when the old speaker produced the conventionalized term. However, when the *old* speaker produced a new term, they found a significant delay in the listener's first look to the target. Listeners in this condition tended to gaze around the display for other objects, suggesting a momentary contrastive implicature. Because the only difference between the two utterances was the identity of the speaker, this provides good reason to believe that listeners also form partner-specific lexical representations over interaction.

1.4.2 Path-dependence and stability across contexts

Another key computational signature of local conventions is their stability, or *stickiness*, across changing contexts. Once a precedent has been established with a particular partner, the pressure to maintain it apparently even trumps the usual Gricean pressures of informativity in context. For example, suppose a pair of participants have converged on an sufficiently specific subordinate term like *pennyloafer* after several rounds referring to a particular shoe in the context of other shoes. When subsequently placed in a new context where all but the target shoe have been replaced with objects from other categories, participants remarkably continued to use the now-overinformative label (e.g. *pennyloafer*) 52% of the time, even though it is the only shoe (Brennan & Clark, 1996).

This can be understood in our model as a consequence of the path-dependence of lexical learning. The initial context affects the initial terms produced via the Gricean reasoning formalized in RSA. Using the basic-level label *shoe* in the initial round would be under-informative because it applies to the distractor shoes equally well (Graf et al., 2016), but there are several appropriately informative alternatives under the lexical prior with approximately the same cost of production (*pennyloafer*, *docksider*, *brown shoe*, *dress shoe*). Which of these roughly equivalent labels the speaker samples, then, is somewhat *arbitrary*: Brennan and Clark (1996) found considerable variance in labeling with only a 10% chance of matching labels across speakers (see also Furnas, Landauer, Gomez, & Dumais, 1987; Hupet et al., 1991).

After a successful round of reference with a particular one of these labels, however, lexica in which this label applies strongly to this particular shoe become more likely, and alternative lexica in which other terms apply better to that same shoe become less likely (by the same Gricean reasoning as earlier: if there were a better term in the partner's lexicon, they would've used it). This dynamic alone accounts for the stability and path-dependence of reference *within* contexts (Hawkins et al., 2017). Empirically, Brennan and Clark (1996) report significantly greater variability of labels across pairs than within pairs (see also Hawkins et al., 2017 for an information theoretic analysis of arbitrariness and stability on a larger data set).

Extending this argument, it also becomes clear why path-dependent learning would stick across new contexts. After several rounds of initial reinforcement, the evidence for the lexical meaning of a subordinate-level term (*pennyloafer*) is so strong that when the context changes, the informativity of this term under the learned lexicon is strong enough relative to the prior uncertainty on the basic-level term (*shoe*) to justify the small additional utterance cost. Critically, this same mechanism also predicts why conventions should *not* generalize when the context switch goes in the opposite direction: when the initial context only contains one shoe, 89% of speakers switched to a more specific utterance when additional shoes are added to the context (Brennan & Clark, 1996). However much speakers strengthened their belief that *shoe* refers to a particular shoe in the first context, it doesn't change the lexical prior of that term *also* applying to the other shoes in the new context. Just saying *shoe* would be extremely underinformative even under the learned lexicon, necessitating a more specific label.

Our lexical learning account also accounts for two additional phenomena related to stability reported by Brennan and Clark (1996): frequency and role independence. First, the more a local convention was initially reinforced, the stickier it was: participants were half as likely to switch to *shoe* from the more specific *pennyloafer* when they did 4 repetitions of *pennyloafer* in the specific context than when they only did 1 repetition. Second, in an experiment when the pair switched roles at the same time as the new context is introduced, this pattern of results stayed the same, indicating that coordinated lexical learning is taking place for both partners.

The stickiness of conventions has been challenged recently by Misyak, Noguchi, and Chater (2016) using a repeated reference game where the communication medium was placing tokens instead of words. They showed that depending on contextual and environmental constraints, the same 'message' (placing a token on a box) can flip its meaning from trial-to-trial in completely contradictory ways. The one-shot pragmatic reasoning allowing for this degree of semantic flexibility presents a fascinating case for lexical uncertainty RSA models, but we strongly disagree that what is being described is a *convention* in any sense discussed in this article via the tradition of Lewis (1969). The severe simplicity of their synchronic scenario (binary signals and binary referents constructed such that there is a single 'optimal' best-response strategy on any trial regardless of prior beliefs) obviates any functional need for diachronic learned representations or dependence on common ground across trials. In particular, a defining property of conventions in is their *arbitrariness*: there needs to exist an alternative that *would* be equally successful if everyone agreed (e.g. participants who use *docksider* are just as successful as those who use *pennyloafer*). There is no such arbitrariness in Misyak et al. (2016), as indicated in their own data showing almost no variability in what participants do on a given trial type. If even the most minimal arbitrariness were introduced, for example two colors of tokens, we predict the usual sticky conventions would form: some participants would begin to persistently associate a particular meaning with 'blue' and others would associate it with 'yellow.'

1.4.3 Generalization from local to global conventions

The first two sections of this paper focused on how global conventions – our lexical priors – shape the formation of local conventions. They seed the initial expectations we bring into interactions, getting communication off the ground and scaffolding rapid lexical learning. But what about influence in the other direction? Where do global conventions come from in the first place? Some are certainly *iconic* and based on expectations about resemblance in shared perceptual systems (Dingemanse, Blasi, Lupyan, Christiansen, & Monaghan, 2015; Verhoef, Kirby, & de Boer, 2016).Yet for the essentially arbitrary form-meaning mappings that make up the bulk of our lexicon, it is difficult to imagine any mechanism for their emergence that doesn't first pass through local conventions. How do interactions with one partner shape the prior one brings into interactions with other partners, and how do these priors converge across social networks?

In Chapter 2, we will propose a hierarchical Bayesian model providing a pathway for such generalization. Hierarchical Bayesian models smoothly integrate population-level expectations with partner-specific ones, and also allow agents to appropriate population-level knowledge through additional partner-specific observations. While we defer the formal specification, it is worth sketching the conceptual framework by analogy to similar models in concept learning. Consider the concept *dog*, which abstracts away from our experiences with different instances of dogs across a lifetime, and provides stable expectations about the properties of a new instance (e.g. four legs, wagging tail, barking noises). However, extensive experience with a particular dog *Fido* also reveals idiosyncratic properties, like their personality or the complex pattern of spots on their coat. If a young (idealized) concept learner has only seen one instance of the category — *Fido* — it may be initially unclear which properties are idiosyncratic and which are shared by all dogs. As they come across more instances, they can draw stronger inferences about what is shared.

Similarly, accumulated knowledge about linguistic conventions in one's community provides

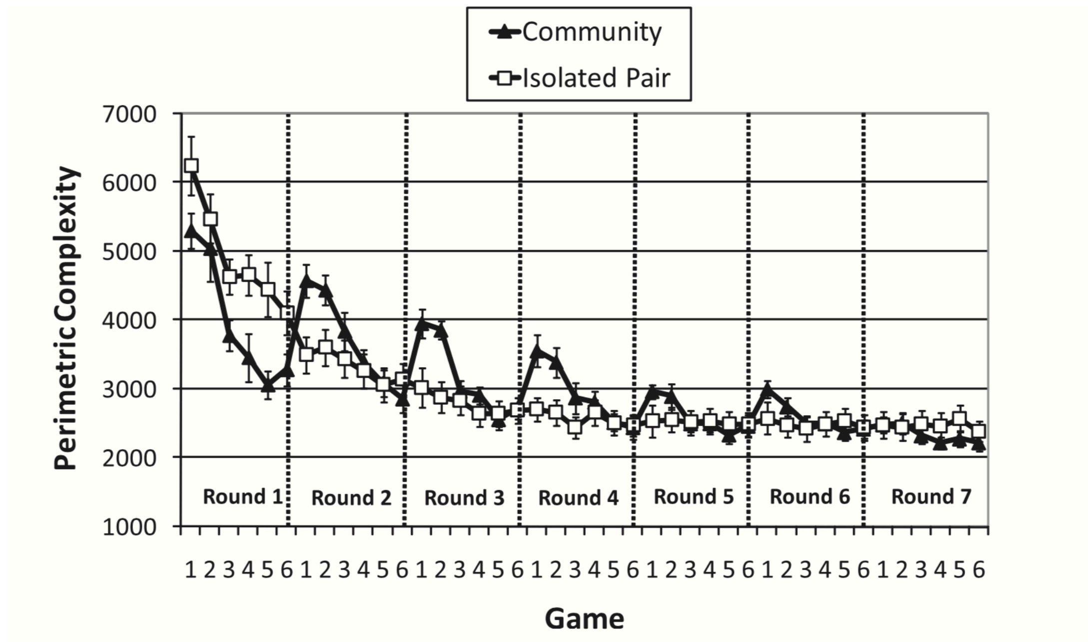

Figure 1.2: Reduction and conventionalization across a community, compared to a single pair. At each vertical line, participants in the "Community" condition switched partners; at each switch, the complexity of their drawings increases, but by less and less as they converge on a shared set of global expectations as efficient as the isolated pair formed over an equal number of games. Reproduced from Fay et al. (2010).

stable communicative priors that guide how we approach new partners. But the language we use to talk with a family member or close collaborator may deviate considerably from the usage predicted by population-level conventions. More than the relatively fixed, biological concept of a dog, though, semantic meanings can be a moving target. The only data we use to ground our learning is produced by other agents who are in the same position as we are, and our only goal is to coordinate on the same meanings in context. So the first time we hear a particular construction (e.g. a slang term or proper noun), it may not be clear whether that partner is just playfully and idiosyncratically making it up, or whether it is shared by others. After we hear several others use the same term in the same way, however, we may be more confident that a novel partner would know it as well.

A simple prediction of a hierarchical learning model is that repeated pairwise interaction within a particular community should lead to convergence within the community as participants begin to generalize across partners. Fay, Garrod, Roberts, and Swoboda (2010) tested this prediction by dividing participants into several 'communities' where they played several rounds of a *graphical* repeated reference game with each member of their community (see below for more discussion of modality differences). Although early partner swaps led to sharp losses in efficiency—consistent with the partner-specificity of conventions—these losses gradually disappeared over successive swaps (see Fig. 1.2), indicating eventual community-level convergence on expectations as strong as those found in isolated pairs (see Garrod & Doherty, 1994, for similar results in a more complex communication game).

A corollary of this in-group convergence is that communication should be *hindered* when assumptions about global priors are violated. When a similar within-group convergence phase was followed by a 'test' phase where participants are either paired with a novel member of their own community or a member of a *different* community (without being explicitly told which it is), sketchers in the latter condition required significantly more ink to succeed. Because no target was referred to more than once with the same partner, whatever alignment of expectations took place over the course of in-group training had to be generalized *across* partners via the overall prior. Note that the network topology implicitly used in these studies (homogenous mixing on a complete graph) is likely critical to this convergence: Centola and Baronchelli (2015) found that simpler coordination games embedded on other common topologies—low-degree lattices and random graphs—tend to get stuck with local regions of the network using incommensurate conventions.

Over longer (generational) time scales, the emergence and stability of community-wide conventions plays a functional role in cultural and linguistic continuity. The composition of a community is constantly shifting as old members die and new members are born: without a mechanism for partner-specific learning to transfer into global beliefs, communication systems would have to be reinvented at each generation. These intergenerational mechanisms have been investigated in the laboratory using replacement or micro-society designs.For instance, Caldwell and Smith (2012) conducted a reference game where a speaker tried to get three listeners to guess color terms by drawing on a sheet of paper. After referring to each color only once, they were replaced by the most senior listener, and a new listener was cycled into the audience. After several rounds, the group was completely different from the original, yet patterns of reduction and path-dependent conventionalization were broadly the same as when a single pair plays repeatedly.

Granted, this particular result could be adequately explained by a non-hierarchical but partner-specific learning model. Because the speaker only had to get *one* of the listeners to guess correctly, they could have relied on the partner-specific models built up for two of their three audience members on previous rounds while simply ignoring the novel partner. But in more extreme cases of pairwise interactions with population turn-over, we nonetheless predict that global conventions would emerge and persist across generations.†

1.5 Discussion

Repeated reference games provide a rich arena for studying social interaction and adaptation. Initial utterances expose the global conventions people bring into novel interactions, and how people incorporate contextual information into their expectations. Successive rounds demonstrate the remarkable speed and flexibility with which people form local conventions to successfully coordinate their behavior. Finally, by intervening in these games to manipulate partner or context, we can reveal the boundaries of learning through tests of generalization.

Throughout, we have argued for a probabilistic model where agents initially have uncertainty over the latent representation guiding their partner's actions and dynamically coordinate over time by conditioning on shared history. A critical component of this model is the hierarchical structure

† Note that actual communication within the lifetime seems crucial for this process: pure iterated learning does not produce conventionalization (Garrod, Fay, Rogers, Walker, & Swoboda, 2010))

by which all members of a community are assumed to be drawn from a distribution with shared parameters, allowing a pathway for global conventions to be influenced by local interactions without sacrificing the ability to learn idiosyncratic partner-specific models. This account situates convention formation as the product of generic hierarchical learning machinery operating on social data. The scope of this review was limited to the broad behavioral phenomena uncovered by repeated reference games, but going forward, it will also be important to (1) tie this computational-level account to underlying neural and algorithmic mechanisms for adaptation and (2) contrast our learning account with both simpler low-level 'priming' based models and richer collaborative notions requiring recursive 'mutual knowledge.'

The speaker wants to be understood. In order to judge how he will be interpreted, he uses his [...] starting theory of interpretation. As speaker and interpreter talk, their "prior" theories become more alike; so do their "passing" theories. Not only does it have its changing list of proper names and gerrymandered vocabulary, but it includes every successful use of any other word or phrase, no matter how far out of the ordinary. Every deviation from ordinary usage, as long as it is agreed on for the moment (knowingly deviant, or not, on one, or both, sides), is in the passing theory as a feature of what the words mean on that occasion. Such meanings, transient though they may be, are literal.

Donald Davidson, 1986

2

An inferential model of convention-formation

In this chapter, we present a computational model formalizing our theoretical account of how speakers coordinate on meaning under uncertainty. Our formal approach is grounded in the family of Rational Speech Act (RSA) models, which have been successful in explaining a wide range of linguistic phenomena—including scalar implicature (N. D. Goodman & Stuhlmüller, 2013), adjectival vagueness (Lassiter & Goodman, 2015), overinformativeness (Degen, Hawkins, Graf, Kreiss, & Goodman, 2019), indirect questions (Hawkins, Stuhlmüller, Degen, & Goodman, 2015), and other non-literal language use (Kao, Wu, Bergen, & Goodman, 2014)—as arising from a process of recursive social reasoning.

At the core of any model of referential communication is the notion of a *semantics* giving the meanings of utterances in the language. For the minimal examples in this chapter, it suffices to define a lexical function $\mathcal{L} : (w, o) \rightarrow \mathbb{R}$, assigning any word-object pair a real-valued meaning according to how well the word w applies to the object o. This is a continuous generalization of classic truth-conditional semantics (Graf et al., 2016), where utterances may apply more or less well to different referents. For instance, the utterance *dancer* may initially be expected to apply to a photorealistic image of a `ballerina` ($\mathcal{L}($*dancer*, `ballerina`$) = 0.99$) more than an abstract sketch of one ($\mathcal{L}($*dancer*, `ballerina drawing`$) = 0.6$), but apply to both better than a non-category member like an image of dog falling down the stairs ($\mathcal{L}($*dancer*, `dog`$) = 0.05$).

In this framework, an nth order pragmatic speaker trying to refer to particular object $o \in \mathcal{O}$ assuming lexicon $\mathcal{L}$ selects an utterance $u \in \mathcal{U}$ by trading off its expected informativity (with respect to a rational listener agent) against its cost, usually based on length:

$$S_n(u|o, \mathcal{L}) \propto \exp\left(\alpha \log L_{n-1}(o|u, \mathcal{L}) - \text{cost}(u)\right)$$

where α is a soft-max optimality parameter controlling the extent to which the speaker maximizes over listener informativity. The listener, in turn, inverts the speaker model to reason about what underlying object o the speaker is trying to convey, given their utterance u:

$$L_n(o|u, \mathcal{L}) \propto S_n(u|o, \mathcal{L}) P(o)$$

This recursion eventually bottoms out in a *literal listener* who directly looks up the meaning of the utterance in the lexicon:

$$L_0(o|u, \mathcal{L}) \propto \mathcal{L}(u, o) \cdot P(o)$$

2.1 Adapting to a single partner

Now, we extend the lexicon from a lookup table or a static logical form into a dynamic, parameterized representation that can be constantly being updated. To formalize this notion of semantic meaning, we begin with the additional assumption of *lexical uncertainty* (Smith et al., 2013; Bergen et al., 2016). That is, instead of assuming agents have fixed knowledge of $\mathcal{L}$, we allow for uncertainty over the exact meanings of lexical items in the current context (e.g. it may be initially unclear what "the dancer" might refer to). Concretely, we put a prior $P(\mathcal{L})$ over the identity of a partner's true lexicon, which may be initially biased toward certain meanings from previous experience.

Bayesian updating then gives a rule for updating expectations about this true lexicon conditioned on repeated observations of a partner's behavior:

$$P_{L_n}(\mathcal{L}|d) \propto P(\mathcal{L}) \prod_i S_n(o_i|u_i, \mathcal{L})$$

where $d = \{o_i, u_i\}$ is a set of observations containing utterances u_i referring to objects o_i from previous exchanges with that partner. The listener marginalizes over this posterior when interpreting a new utterance u from the speaker:

$$L_n(o|u, d) \propto \sum_{\mathcal{L}} P_{L_n}(\mathcal{L}|d) L_n(o|u, \mathcal{L})$$

The speaker, in turn, considers what utterances would be most informative for such a listener:

$$S_n(u|o, d) \propto \exp(\alpha \log \left(\sum_{\mathcal{L}} P_{S_n}(\mathcal{L}|d) L_{n-1}(o|u, \mathcal{L}) \right) - \text{cost}(u))$$

where the posterior over lexica $P_{S_n}(\mathcal{L}|d)$, uses the listener likelihood L_{n-1} instead of the speaker

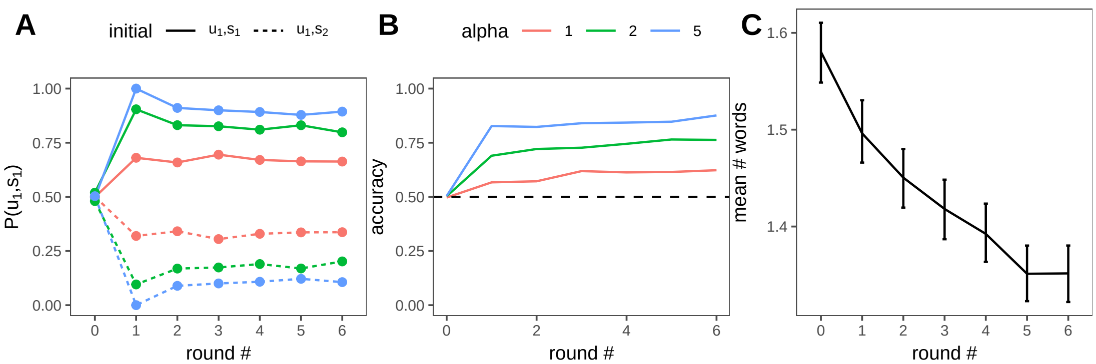

Figure 2.1: Schematic of model

likelihood:

$$P_{S_n}(\mathcal{L}|d) \propto P(\mathcal{L}) \prod_i L_{n-1}(u_i|o_i, \mathcal{L})$$

For the purposes of this paper, we fix the depth of recursion at $n = 2$. This model is implemented in the probabilistic programming language WebPPL (N. D. Goodman & Stuhlmüller, electronic). *.

2.1.1 Model Results

Coordination First, we show how agents updating their meaning functions in this way can coordinate even in the absence of strong initial priors. The initial choices in an interaction can be taken as evidence for a particular lexicon and become the basis for successful communication, even when both speaker and listener are uncertain at the beginning. As a simple test case, consider an environment with two objects ($\{o_1, o_2\}$), where the speaker must choose between two utterances ($\{u_1, u_2\}$) with equal production costs. For the prior $P(\mathcal{L})$ over the meaning of each utterance, we define a Beta distribution†, so on the first round both utterances are equally likely to apply to either shape. If the speaker were trying to get their partner to pick o_1, then because each utterance is

*These results can be reproduced running our code in the browser at `http://forestdb.org/models/conventions.html`

†In our implementation, we use exact enumeration over coarse-grained bins; experiments using variational inference on the full continuous distribution give similar results

equally (un)informative, they could only randomly sample one (say, u_1), and observe the listener's selection of a shape (say, o_1, a correct response). On the next round, the speaker uses the observed pair $\{u_1, o_1\}$ to update their beliefs about their partner's true lexicon, uses these beliefs to generate a new utterance, and so on. To examine expected dynamics over multiple rounds, we forward sample many possible trajectories.

We observe several important qualitative effects in our simulations. First, and more fundamentally, the evidence that a knowledgeable listener responded to utterance u by choosing a particular object o provides support for lexicons in which u is a good fit for o. Hence, the likelihood of the speaker using u to refer to o will increase on subsequent rounds (see Fig.2.1A). In other words, the initial symmetry between the meanings can be broken by initial random choices, leading to completely arbitrary but stable mappings in future rounds.

Second, because the listener is updating their meaning representation from the same observations under the same set of assumptions, both partners converge on a *shared* set of meanings; hence, the expected accuracy of selecting the target object rises on future rounds (see Fig. 2.1B). Third, because one's partner is assumed to be pragmatic via recursive Rational Speech Act mechanisms, agents can also learn about *unheard* utterances. Observing $d = \{(u_1, o_1)\}$ also provides evidence that u_2 is *not* a good fit for o_1. This effect arises from Gricean maxims: if u_2 were a better fit for o_1, the speaker would have used it instead (Grice, 1975). Fourth, *failed references* can lead to conventions just as effectively as successful references: if the speaker intends o_1 and says u_1, but then the listener incorrectly picks o_2, the speaker will take this as evidence that u_1 actually means o_2 in their partner's lexicon and become increasingly likely to use it that way on subsequent rounds.

Reduction Next, we show how our model explains reduction of utterance length over multiple interactions. For utterances to be reduced, of course, they must vary in length, so we extend our grammar to include *conjunctions*. Conjunctions are one of the simplest ways to constructi longer utterances compositionally from lexical primitives, using the product rule:

$$\mathcal{L}(u_i \text{ and } u_j, o) = \mathcal{L}(u_i, o) \times \mathcal{L}(u_j, o)$$

Analogous to the *tangram* stimuli used in the reference game reviewed in Chapter 1, which have many ambiguous features and figurative perspectives that may be evoked in speaker descriptions, we consider a scenario the two objects $\{o_1, o_2\}$ differ along two different features. The speaker thus has four primitive words at their disposal – two words for the first feature ($\{u_{11}, u_{12}\}$) and two for the second $\{u_{21}, u_{22}\}$. While we established in the previous section that conventions can emerge over a reference game in the complete absence of initial preferences, players often bring such preferences to the table. A player who hears 'ice skater' on the first round of a tangrams task is more likely to select some objects more than others, even though they still have some uncertainty over its meaning in the context. To show that our model can accommodate this fact, we allow the speaker's initial prior meanings to be slightly biased. We assume u_{11} and u_{21} are a priori more likely to mean o_1 and u_{12} and u_{22} are more likely to mean o_2.

We ran 1000 forward samples of 6 rounds of speaker-listener interaction, and averaged over the utterance length at each round [‡]. Our results are shown in Figure 2.1C: the expected utterance length decreases systematically over each round. To illustrate in more detail how this dynamic is driven by an initial rational preference for redundancy relaxing as reference becomes more reliable, we walk step-by-step through a single trajectory. Consider a speaker who wants to refer to object o_1. They believe their knowledgeable partner is slightly more likely to interpret their language using a lexicon in which u_{11} and u_{12} apply to this object, due to their initial bias. However, there is still a reasonable chance that one or the other alone actually refers more strongly to o_2 in the true lexicon. Thus, it is useful to produce the conjunction "u_{11} and u_{12}" to hedge against this possibility, despite its

[‡]In our simulations, we used $\alpha = 10$ and found the basic reduction effect over a range of different biases

higher cost. Upon observing the listener's response (say, o_1), the evidence is indeterminate about the separate meanings of u_{11} and u_{12} but both become increasingly likely to refer to o_1. In the trade-off between informativity and cost, the shorter utterances remain probable options. Once the speaker chooses one of them, the symmetry collapses and that utterance remains most probable in future rounds. In this way, meaningful sub-phrases are omitted over time as the speaker becomes more confident about the true lexicon.

2.2 Generalizing to new partners

The model introduced in the previous section formally combines the first two mechanisms outlined in Chapter 1: a lexical prior and a learning mechanism to rapidly coordinate with a particular partner. However, it leaves the question of *generalization* unaddressed: what happens when we start talking to a new partner? To what extent, if any, does an interaction with one partner change our expectations about the next? In this section, we develop a theory of generalization capturing how agents transfer what they have learned about one partner to another partner.

Several theories of generalization have been proposed in prior work, but each has been limited in different ways. One possible theory is that speakers simply collapse over the identity of different partners, using the posterior from one partner as the prior for the next. This idea is common among accounts of cultural evolution that draw on network science and economic games. By using a single (egocentric) representation of linguistic conventions irrespective of one's partner, these accounts explain the emergence of conventions at the population level over long time scales via distributed interactions among minimal agents (Steels, 1995; Barr, 2004; Young, 2015; Spike, Stadler, Kirby, & Smith, 2017; Baronchelli, 2018; Hawkins, Goodman, & Goldstone, 2019).

A second theory is that speakers maintain separate representations of the meanings shared with each social partner. In the simplest versions of this theory, information built up through the distinct history of interaction with one partner should crucially *not* generalize to other partners (H. H. Clark, 1996; Brennan & Hanna, 2009; Horton & Gerrig, 2016). This idea has been fundamental for explaining the flexibility with which speakers adapt and switch between different social partners (H. H. Clark & Wilkes-Gibbs, 1986; Metzing & Brennan, 2003) over the shorter timescales of language use studied in psycholinguistics.

These two theories of generalization—ignoring partner identity versus maintaining distinct partner-specific representations—mirror the contrast between *complete pooling* and *no pooling* in multi-level statistical models (Gelman & Hill, 2006). Like their statistical counterparts, these two theories of generalization in language face problems: each one is unable to explain the characteristic phenomena motivating the other. Ignoring partner identity will fail in the settings of interest in psycholinguistics where distinct information about one's partner is available and relevant for conversation. On the other hand, entirely sealing off the meanings built up with one partner from expectations about others means speakers will revert to a fixed backdrop with each novel partner[§].

Here, we propose a third formal alternative, inspired by the solution to similar problems in the statistics and machine learning literature. Hierarchical *partial pooling* models offer a compromise between these extremes: agents equipped with such a model can integrate rich knowledge of specific partners with a mechanism for abstracting away what is common across different partners. These models have been key to explaining how the human mind solves other difficult inductive problems in domains like causal learning (Kemp, Goodman, & Tenenbaum, 2010; N. D. Goodman, Ullman, & Tenenbaum, 2011), speech perception (Kleinschmidt & Jaeger, 2015) and concept learning (Kemp, Perfors, & Tenenbaum, 2007) where abstract, shared properties must be jointly inferred with idiosyncratic particulars of instances. Such a model has the potential to span the gap between these

[§]Of course, theoretical surveys have extensively noted the importance of *cultural* common ground associated with entire social groups rather than particular individuals (H. H. Clark, 1998, e.g.). This distinction implicitly creates an avenue for generalization to new partners, insofar as talking to one individual is informative about that broader cultural common ground. Our approach formalizes this general idea.

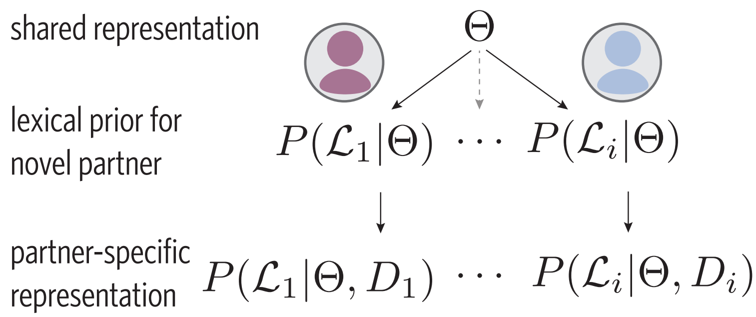

Figure 2.2: Schematic of hierarchical model

different traditions, and to capture how agents make use of multiple levels of social structure across partners and communities.

We formalize this model as a straightforward extension of the RSA model above. Specifically, we add an additional community-level representation parameterizing the lexical prior to create a hierarchical dependency structure (see Fig. 2.2). At the highest level of the hierarchical lexical representation is a *community-level* variable Θ parameterizing the agent's prior expectations for the likely lexicon $\mathcal{L}_i$ used by a novel community member i: $P(\mathcal{L}_i|\Theta)$. For simplicity, we generalize our Beta prior in the earlier section by assuming Θ is a tensor parameterizing a matrix containing an entry for every (w, o) in the lexicon $\mathcal{L}_i$. We can then place an uninformative prior $P(\Theta)$ over that tensor which does not overwhelm the likelihood. We will use a diagonalized multivariate Gaussian (but see Gelman et al., 2014, p. 110, for other reasonable choices). More generally, we could allow for arbitrarily complex dependencies between entries of the lexicon by using a neural network with weight tensor Θ. In Chapter 4, we will implement such a prior using an Empirical Bayes approach, only representing a point estimate of Θ rather than the whole distribution. For now, it only matters that this knowledge is hierarchical: we expect all members of our language community to share some commonality in what they mean by things.

Now that we have defined a hierarchical likelihood on lexical beliefs, we must say how we *learn* partner-specific models. Our partner-specific beliefs about a particular individual's semantics $\mathcal{L}_i$ are formed by integrating our abstract lexical knowledge Θ with particular observations D_i of that particular individual, concretely, utterances and responses in a reference game:

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

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

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

Note that our posterior beliefs about Θ are in fact informed by observations from *all* speakers: $D = \bigcup_{i=1}^k D_i$. Additionally, because the partner-specific model depends on Θ, Bayesian inference allows new data to systematically inform the shared, population-level representation as well (Fig. 2.2). Critically for predictions about generalization, new language data (i.e. particular ways of referring to the tangram shapes) may at first be more parsimoniously explained as an idiosyncratic property of a particular partner's lexicon, or "idiolect". If two or three partners all happen to use the same language, however, it starts to become more likely that a novel partner will share it as well (this transfer is sometimes referred to as "sharing of statistical strength.") This formalizes the intuition sketched in Chapter 1.

Finally, to fully specify our model and compute our partner-specific lexical posterior $P(\mathcal{L}_i|D_i, \Theta)$, we must link our beliefs about a partner's lexica to their actual behavior with a likelihood function $P(D_i|\mathcal{L}_i, \Theta)$. This is naturally supplied by the Rational Speech Act framework used in the previous section (Frank & Goodman, 2012; N. D. Goodman & Frank, 2016; Bergen et al., 2016; Smith et al., 2013): 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 as well as a cue to the communal lexicon shared.

In summary, our hierarchical model formalizes the intuition that global conventions are learned and generalized over repeated interactions with many different people, and that this shared semantic prototype is the backbone supporting rapid learning for new partners and situations.

Model Results Our primary aim in this section is to examine the gradient of generalization produced by our hierarchical model across different partners. To test this behavior, we used the same simple scenario with two objects and two words as we used to test adaptation to a single partner. As our measure of interest, we focus on a listener agent's shifting expectations about which target is being referred to as they adapt their semantic representations. Instead of presenting a stream of data from a single partner, however, we swap in a new partner every k rounds of the reference game.

To derive quantitive behavior in this scenario, we must concretely specify the agent's lexical priors and a method to perform inference. We used independent Gaussian distributions for each $\theta_{ij} \in \Theta$ as a hyper-prior and centered our partner-specific prior $\ell_{ij} \in \mathcal{L}$ at the shared value for a particular partner:

$$\begin{aligned} P(\theta_{ij}) &\sim \mathcal{N}(0, 5) \\ P(\ell_{ij}) &\sim \mathcal{N}(\theta_{ij}, 1) \end{aligned}$$

The variances chosen in these priors are assumptions about how strongly adaptation is regularized. Using a higher variance for the hyper-prior than for the partner-specific prior corresponds to the assumption that the agent has substantial uncertainty over the population-level lexicon, but that individuals in the population shouldn't vary too much from whatever that lexicon is.

We used a variational method (VI; Ranganath, Gerrish, & Blei, 2013; Blei, Kucukelbir, & McAuliffe,

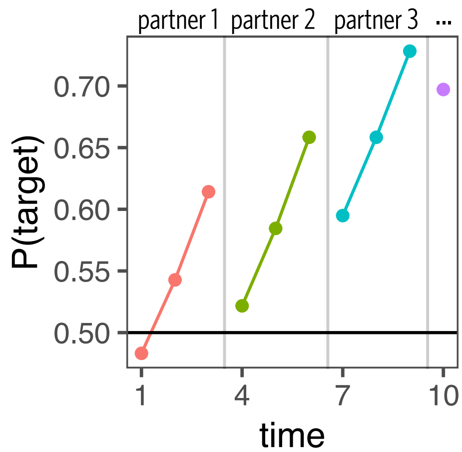

Figure 2.3: Listener predictions as more evidence accumulated from different partners.

2017) to perform inference in our model. Variational methods transform probabilistic inference problems into optimization problems by approximating the true posterior with a parameterized family. Specifically, we make a *mean-field* approximation and assume that the full posterior can be factored into independent Gaussians for each random variable in the lexicon. We then optimize the parameters of these posterior Gaussians by minimizing the evidence lower bound (ELBO) objective (see Murphy, 2012, for more details). To examine how inferences change over time, we amortize inference. We run 10,000 steps of gradient descent on the first observation to obtain a lexical posterior, compute the listener's marginal prediction for the next observation, then continue running gradient descent on the same stored parameters after including the new observation in the data.

Our simulation results are shown in Fig. 2.3. To observe behavior in the simplest case, we feed the model the same object-target pair ($\{o_1, w_1\}$) on every trial. We find that the adapting listener begins at chance because of its uninformative prior, but after interacting with that partner for several rounds, it adapts its lexicon to select the target well above chance. When a new partner is introduced, however, it reverts nearly to its original state: because of the hierarchical structure of its lexical expectations, it was ambiguous whether the evidence from the first partner was idiosyncratic or due to shared structure. After adapting similarly to the second partner, however, it begins its interaction with a third partner with nearly the accuracy at the end of its interaction with the first partner. Thus, its expectations about individual partners' lexica are gradually being incorporated into community-level expectations.

3

Continuous adaptation for efficient machine communication

Linguistic communication depends critically on shared knowledge about the meanings of words (Lewis, 1969). However, the real-world demands of communication often require speakers and listeners to go *beyond* dictionary meanings to understand one another (H. H. Clark, 1996; Hasson, Ghazanfar, Galantucci, Garrod, & Keysers, 2012; Stolk, Verhagen, & Toni, 2016). The social world continually presents new communicative challenges, and agents must continually coordinate on new meanings to meet them. For example, consider a nurse visiting a bed-ridden patient in a cluttered home. The first time they ask the nurse to retrieve a particular medication, the patient must

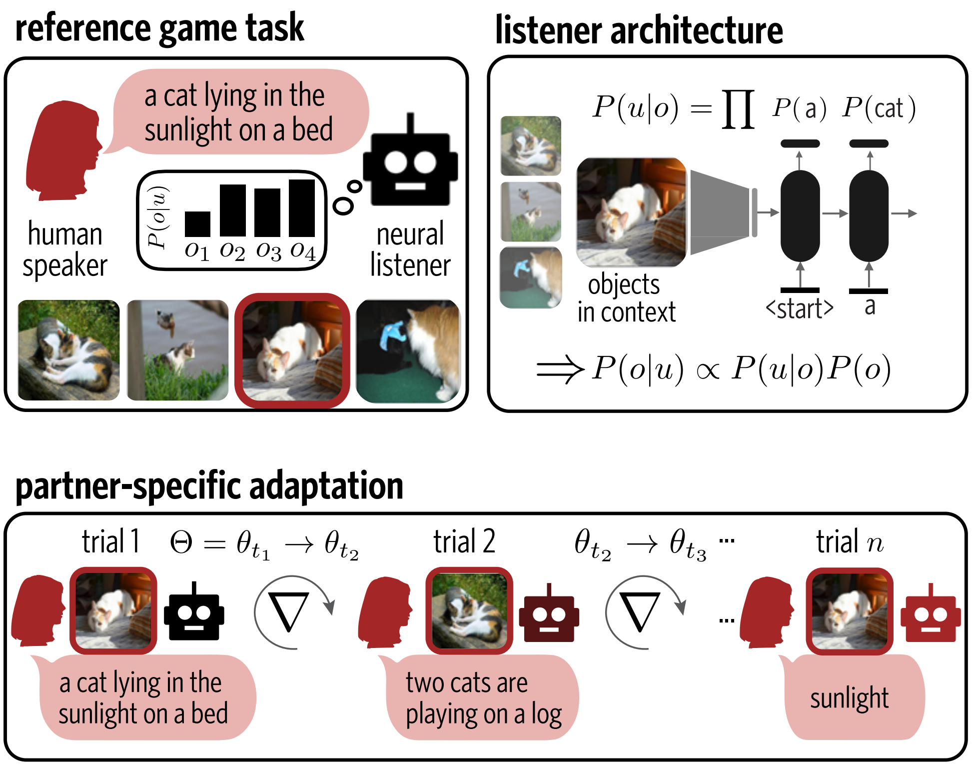

Figure 3.1: Reference game task, listener architecture, and continual learning approach.

painstakingly refer to unfamiliar pills, e.g. "my vasoprex-tecnoblek meds for blood pressure, in a small bluish bottle, on the bookcase in my bathroom." After a week of care, however, they may just ask for their "Vasotec."

This type of flexible language use poses a challenge for models of language in machine learning. Approaches based on deep neural networks typically learn a monolithic meaning function during training, with fixed weights during use. For an in-home robot to communicate as flexibly and efficiently with patients as a human nurse, it must be equipped with a continual learning mechanism. Such a mechanism would present two specific advantages for interaction and communication applications. First, to the extent that current models have difficulty communicating in a new setting, an adaptive approach can quickly improve performance on the relevant subset of language. Second, for human-robot contexts, an adaptive model enables speakers to communicate more efficiently as they build up common ground, remaining understandable while expending significantly fewer words as humans naturally do (H. H. Clark & Wilkes-Gibbs, 1986).

In this chapter, we introduce a general framework for transforming neural language models into *adaptive* models that can be deployed in real-time interactions with other agents. Our key insight is that through continual interactions with the same partner in a shared context, a listener can adapt and more efficiently communicate with its partner (Fig. 3.1). We are motivated by the hierarchical Bayesian approaches to task-specific adaptation from Chapter 2. We implement this theoretical approach at scale by integrating two core components: (i) a loss function combining speaker and listener information to understand descriptions of natural images in context, and (ii) a regularization scheme for fine-tuning the weights of this model based on previous interactions with a partner. We show that these components enable more effective communication with human partners over repeated interactions.

3.1 Approach

We begin by recasting communication as a multi-task problem for meta-learning. Each context and communicative partner can be regarded as a related but distinct task making its own demands on the agent's language model. To be effective across many such tasks, a communicative agent must both (1) have a good prior representation they can use to understand novel partners and contexts, and (2) have a mechanism to rapidly update this representation from a small number of interactions.

3.1.1 Repeated reference game task

As a benchmark for studying this problem, we introduce the *repeated reference game* task (Fig. 3.1), which has been widely used in cognitive science to study partner-specific adaptation in communication (Krauss & Weinheimer, 1964; H. H. Clark & Wilkes-Gibbs, 1986; Wilkes-Gibbs & Clark, 1992). This task is a special case of the more general family of reference games, where a speaker agent is given a context of N images (a target object o among $N-1$ distractors) and must produce an utterance u that allows their partner, a listener agent, to identify o with high probability. In a *repeated reference game*, each image in context appears as the target multiple times, allowing us to evaluate how communication about a particular image changes as the speaker and listener build up a shared history.

3.1.2 Continual adaptation with Hierarchical Bayes

Before formalizing our algorithm as a generic update rule for neural networks, we describe the theoretical Bayesian foundations of our approach, reviewing material from previous chapters in the context of machine learning. Under a Bayesian approach, semantic representations can be viewed probabilistically: we represent some uncertainty over meanings. In a hierarchical Bayesian model, this uncertainty is structured over different partners and contexts.

At the highest level is a *task-general* variable Θ which parameterizes the agent's task-specific prior expectations $P(\theta_i|\Theta)$ where θ_i represents the semantics used by a novel partner i. Given observations D_i from communicative interactions in that context, an agent can update their *task-specific* model using Bayes rule:

$$P(\theta_i|D_i,\Theta) \propto P(D_i|\theta_i)P(\theta_i|\Theta) \tag{3.1}$$

The Bayesian formulation thus decomposes the problem of task-specific adaptation into two terms, a prior term $P(\theta_i|\Theta)$ and a likelihood term $P(D_i|\theta_i)$. The prior captures the idea that different language tasks share some task-general structure in common: in the absence of strong information about usage departing from this common structure, the agent ought to be regularized toward their task-general knowledge.

The likelihood term accounts for needed deviations from general knowledge due to evidence from the current situation. The form of the likelihood depends on the task at hand. For the referential communication task we consider here, $D_i = \{(u, o)_t\}$ contains paired observations of utterances u and their objects of reference o at time t. These data can be viewed from the point of view of a speaker (generating u given o) or a listener (choosing o from a context of options, given u) (Smith et al., 2013; Hawkins et al., 2017); these yield different likelihoods that update the semantics in complementary ways. The generative model of a *speaker* uses the task-specific semantics θ_i to sample utterances u proportional to how well they apply to o:

$$P_S(u|o, \theta_i) \propto \exp f_{\theta_i}(u, o) \tag{3.2}$$

A *listener* can be modeled as inverting this speaker model to evaluate how well an utterance u describes each object o *relative to the others* in a context $\mathcal{C}$ of objects by normalizing (Frank & Goodman, 2012; Vedantam, Bengio, Murphy, Parikh, & Chechik, 2017; Cohn-Gordon, Goodman, & Potts, 2018; Monroe, Hawkins, Goodman, & Potts, 2017):

$$P_L(o|u, \mathcal{C}, \theta_i) \propto P_S(u|o, \theta_i)P(o) \tag{3.3}$$

Because these views of the data from past interactions D_i provide complementary statistical information about the task-specific semantics θ_i, we will combine them in our loss.

3.1.3 Continual adaptation for neural language models

While the $N \times M$ word-object matrix used as the semantics in Chapter 2 cannot straightforwardly generalize to unseen words and objects. It also quickly becomes intractable as the vocabulary and object set grows, making it untenable for modeling arbitrary natural language. Here we will instead take Θ to be an initialization for the weights of an image-captioning neural network (see Fig. ??A).

While it is theoretically possible to place priors on all neural network parameters and attempt to jointly approximate posteriors (Joshi et al., 2017), current techniques for inference in Bayesian neural networks rely on optimization of noisy gradients of variational objectives and tend not to work well. Instead, because maintaining full hyper-priors is costly and challenging for inference, we will assume agents only represent an *empirical Bayes* point estimate of Θ (Gelman et al., 2014). Additionally, we exploit a deep theoretical connection between the hierarchical Bayesian framework presented in the previous section and recent deep learning approaches to multi-task learning (Nagabandi, Finn, & Levine, 2018; Grant, Finn, Levine, Darrell, & Griffiths, 2018; Jerfel, Grant, Griffiths, & Heller, 2018). Given a task-general initialization, regularized gradient descent on a particular task is equivalent to conditioning on new data under a Bayesian prior. We exploit this connection to propose an online continual learning scheme for a neural listener model that can adapt to a human speaker in a challenging referential communication task.

Concretely, we consider an image-captioning network that combines a convolutional visual encoder (ResNet-152) with an LSTM decoder (Vinyals, Toshev, Bengio, & Erhan, 2015). The LSTM takes a 300-dimensional embedding as input for each word in an utterance and then uses a softmax layer to linearly project back to a distribution over the vocabulary size. An adapter replacing the final fully connected layer of the encoder was jointly pre-trained with the decoder on the COCO training captions and then frozen as our task-general initialization Θ. For each utterance-object data point observed in the current task, we take a small number of gradient steps fine-tuning the decoder's weights to better account for the speaker's usage (see Algorithm 1). We consider several loss terms and techniques to do so.

Speaker and listener likelihood. The primary signal available for adaptation is the (log-) probability of the new data under speaker and listener likelihoods given in Eqns. 3.2-3.3. Our speaker

Algorithm 1 : Update step for adaptive language model

Input: θ_t: weights at time t
Output: θ_{t+1}: updated weights
Data: (u_t, o_t): observed utterance and object at time t
for step do
  sample augmented batch of sub-utterances u $\sim \mathcal{P}(u)$
  update $\theta_t \leftarrow \theta_t + \beta\nabla[P(\mathrm{u}|o) + P(o|\mathrm{u}) + \mathrm{reg}(o, u)]$
end for

likelihood serves to make the observed utterance more likely for the target in *isolation*, while our listener likelihood makes it more likely *relative* to other objects in context. The speaker and listener likelihoods can be computed directly from the neural captioning model, where each word is distributed according to a softmax over the LSTM output given the sentence so far.

Regularization. We introduce three kinds of regularization terms to approximate the Bayesian prior on task-specific learning. First, rather than directly regularizing weights, a *speaker KL regularization* term minimizes the divergence between the captioning model's output probabilities before and after fine-tuning (Yu, Yao, Su, Li, & Seide, 2013; Galashov et al., 2018). Since the support for our distribution of captions are infinite, we approximate the divergence incrementally by expanding from the maximum a posteriori (MAP) word at each step according to P, where P represents the model at initialization and Q_t represents the model at time t. This loss is then averaged across random images from the full domain $\mathcal{O}$, not just those in context:

$$\sum_{o\sim\mathcal{O}}\sum_{i} D_{\mathrm{KL}}\left(P(w_i|o, w_{i-1}^{\mathrm{MAP}})||Q_t(w_i|o, w_{i-1}^{\mathrm{MAP}})\right) \tag{3.4}$$

Second, we derive a *listener KL regularization* term which compares the initial listener distribution over objects in context $o \in \mathcal{C}$ with the fine-tuned model's distribution: $D_{\mathrm{KL}}\left(P(o|u)||Q_t(o|u)\right)$. The third form of regularization we consider is *local rehearsal*. We evaluate our listener likelihood over prior observations $(u, o) \in D_i$ to prevent overfitting to the most recent observation. To capture how the likelihood overwhelms the prior with increasing data in Bayesian updating, we anneal the listener regularization and rehearsal over the course of interaction while reverse-annealing the listener likelihood.

Data augmentation. A final component of our algorithm is the introduction of a data augmentation step on the new utterance u. Ideally, an adaptive agent should learn that sub-components of the observed utterance are compositionally responsible for this meaning. We thus derive a small training dataset $D(u)$ from u; for simplicity, we take the (ordered) powerset $D(u) = \mathcal{P}(u)$ of all sub-utterances.*

3.2 Evaluations

To evaluate our model, we implemented a repeated reference game using images from the validation set of COCO (Lin et al., 2014) as the targets of reference. To construct challenging contexts $\mathcal{C}$, we used our pre-trained visual encoder to find sets of highly similar images. We extracted feature vectors for each image, partitioned the images into 100 groups using a k-means algorithm, sampled one image from each cluster, and took its 3 nearest neighbors in feature space, yielding 100 unique contexts of 4 images each†.

3.2.1 Human baselines

We first investigated the baseline performance of human speakers and listeners. We recruited 113 participants from Amazon Mechanical Turk and automatically paired them into an interactive environment with a chatbox. For each of these 56 pairs, we sampled a context and constructed a sequence of

*Grammatical acceptability could in principle be taken into account using alternative sets derived from a syntactic parse.

†Using pre-trained VGG as the encoder gave qualitatively similar contexts.

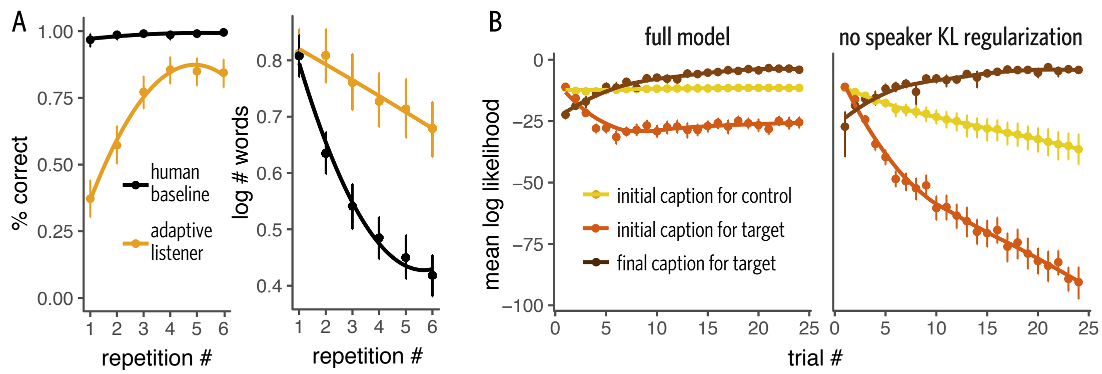

Figure 3.2: (A) Human speakers grow more efficient and accurate as our model adapts. Curves show regression fits. (B) Speaker KL regularization prevents catastrophic forgetting. Error bars and ribbons are bootstrapped 95% CIs.

24 trials structured into 6 repetition blocks, where each of the 4 images appeared as the target once per block. We prevented the same target appearing twice in a row and scrambled the order of the images on each player's screen on each trial.

We found that pairs of humans were remarkably accurate at this task, with performance near ceiling on every round. At the same time, they grew increasingly efficient in their communication: the utterance length decreased from an average of 7 words per image on the first repetition to only 3 words on the last. A mixed-effects regression with random slopes and intercepts accounting for variability at the pair- and context-level found a significant decrease in utterance length across repetitions, $t = -5.8, p < 0.001$ (Fig. 3.3A).

3.2.2 Model evaluation with human partner

Next, we evaluated how our adaptive listener performed in *real-time interaction* with human speakers. We recruited 45 additional participants from Amazon Mechanical Turk who were told they would be paired with an artificial agent learning how they talk. This task was identical to the one performed by humans, except participants were only allowed to enter a single message through the chatbox on each trial. This message was then 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. The

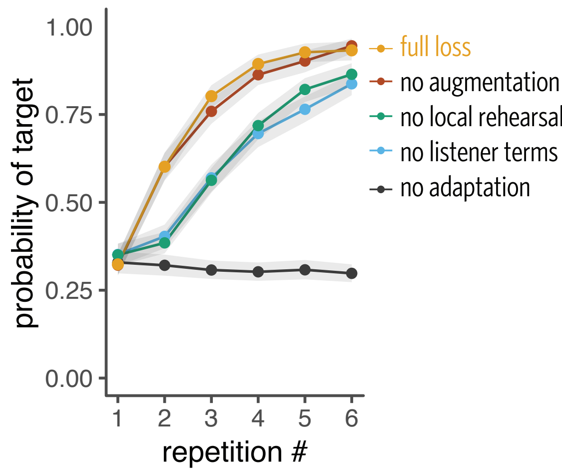

Figure 3.3: Lesions reveal the contributions of each loss term. Error bars and ribbons are bootstrapped 95% CIs.

approximate latency for the model to respond was 6-8 seconds.

We used a batch size of 8, learning rate of 0.0005, and took 8 gradient steps after each trial. For our loss objective, we used a linear combination of the speaker likelihood loss, listener likelihood loss, and all three regularization terms. We found that a listener based on a pre-trained neural captioning model—the initialization for our adapting model—performs much less accurately than humans due to the challenging nature of the reference task. Yet our model rapidly improves in accuracy as it coordinates on appropriate meanings with human speakers. Similarly, while speakers did not simplify their utterances to the same extent as they did with other humans, perhaps due to early feedback about errors, they nonetheless became significantly more efficient over time, $b = -19, t = -5$ (see Fig. 3.3A).

3.3 Analysis

We proceed to a series of lesion analyses that analyze the role played by each component of our approach.

3.3.1 Preventing catastrophic forgetting

Fine-tuning repeatedly on a small number of data points presents a clear risk of catastrophic forgetting (Robins, 1995), losing our ability to produce utterances for other images. Our speaker KL regularization (Eqn. 3.4) was intended to play the same role as a Bayesian prior, preventing catastrophic forgetting by tethering task-specific behavior to the task-general model. To test the effectiveness of this term, we examined the likelihood of different captions before and after adaptation to the human baseline utterances. First, we sampled a random set of images from COCO that were not used in our experiment as *control* images, and used the initialized state of the LSTM to greedily generate a caption for each. We also generated initial captions for the *target* objects in context. We recorded the likelihood of all of these sampled captions under the model at the beginning and at each step of adaptation until the final round. Finally, we greedily generated an utterance for each target at the end and retrospectively evaluated its likelihood at earlier states. These likelihood curves are shown with and without speaker KL regularization in Fig. 3.3B. The final caption becomes more likely in both cases; without the KL term, the initial captions for both targets and unrelated controls are (catastrophically) lost.

3.3.2 Lesioning loss terms

We next simulated our adaptive listener's performance hearing utterances from the human baseline under lesioned losses (Fig. 3.3C). We found that rehearsal on previous rounds had the largest qualitative benefit, allowing for faster adaptation on early rounds, while data augmentation and the non-rehearsal listener terms provided small boosts later in the game. Compared to a non-adapting baseline, however, even a simple loss only containing the speaker likelihood and speaker KL regularization performed better over time—successfully adapting to human language use.

HAMLET: Do you see yonder cloud that's almost in shape of a camel?
POLONIUS: By th' mass, and 'tis like a camel indeed.
HAMLET: Methinks it is like a weasel.
POLONIUS: It is backed like a weasel.
HAMLET: Or like a whale.
POLONIUS: Very like a whale.

Shakespeare – Hamlet, Act 3, Scene 2

4 Characterizing conventions: the dynamics of structure and content

Human language use is flexible, continuously adapting to the needs of the current situation. In the previous chapter, we introduced a challenging repeated reference game benchmark for artificial agents, which requires such adaptability to succeed. Inspired by the Bayesian model in Chapter 2, we proposed a continual learning approach that forms context-specific conventions by adapting general-purpose semantic knowledge. Even when models based on general-purpose knowledge perform poorly, our approach allowed human speakers working with adapted variants of such models to become more accurate and more efficient over time.

The computational approach developed in these chapters explains key qualitative features of how speakers and listeners may dynamically solve the problem of talking with new partners about new referents. However, further model development depends critically upon a finer-grained characterization of the *quantitative* signatures of semantic adaptation found in human communication. Certain fundamental descriptive questions remain unanswered, and important theoretical constructs remain poorly operationalized. For example, it has been widely observed that utterances reduce in length as common ground is accumulated. But a precise characterization *what* gets reduced, and *how*, has remained elusive. What determines whether a particular word is dropped or preserved? Are words dropped randomly or systematically in phrases? Similarly, while theoretical definitions of constructs like arbitrariness or stability have loomed over the theoretical analysis of conventions (Lewis, 1969), it has been unclear how exactly to measure the extent to which these properties hold in a particular task and how they may evolve over the course of interaction. Without addressing these gaps in measurement, it is difficult to set criteria to distinguish among different models.

In this chapter, we examine these questions in a large corpus of referring expressions from a new web-based replication of the classic Tangrams task (H. H. Clark & Wilkes-Gibbs, 1986). The computational techniques necessary to analyze such rich natural language data were limited at the time of prior work, but have become newly tractable given developments in natural language processing (NLP). Our analyses divide into two broad categories roughly corresponding the dynamics of *content* and *structure* of referring expressions across interaction. To examine content, we extracted word embeddings (e.g. GloVe vectors) for each message to calculate the similarity of messages within and across pairs. We found that while different pairs coordinate on a wide range of idiosyncratic solutions to the problem of reference, they do so in an increasingly stable and path-dependent manner. Further, 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. To examine structure, we extracted parts of speech and syntax trees from the text to understand what was reducing and how.

We 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. These findings provide higher resolution into the quantitative dynamics of convention formation and support the modeling framework introduced in earlier chapters. Based on usage, new meanings are systematically grounded with a partner to support more efficient communication.

4.1 Methods: Repeated reference experiment

To collect a large corpus of natural dialogue that allows us to measure how pairs coordinate on meaning over time, we faced two primary decisions. First, to observe the formative period of linguistic conventions, we required novel, ambiguous stimuli for which participants didn't already have strong initial conventions. Second, to observe the *dynamics* of conventions over time, we needed the same coordination problem to be repeated over time, such that earlier outcomes are relevant for later decisions. These criteria are satisfied by a *repeated reference game* design in which participants refer to the same objects across multiple rounds as they build up a shared history of interaction, or common ground, with their partner.

We developed two variants of the game: a relatively unconstrained *free-matching* version that more closely replicates the classic in-lab design, and a more tightly controlled *cued* version that allows for higher resolution analyses of how references to individual tangrams changed over time (see Fig. 4.1). The *free-matching* version was an exploratory sample, but we pre-registered our full preprocessing and analysis pipeline for the *cued* version[*]. While we have released the corpus from the free-matching version[†], we privilege the *cued* version throughout the paper as our confirmatory sample.

[*] https://osf.io/2zwmx

[†] https://cocolab.stanford.edu/datasets/tangrams.html

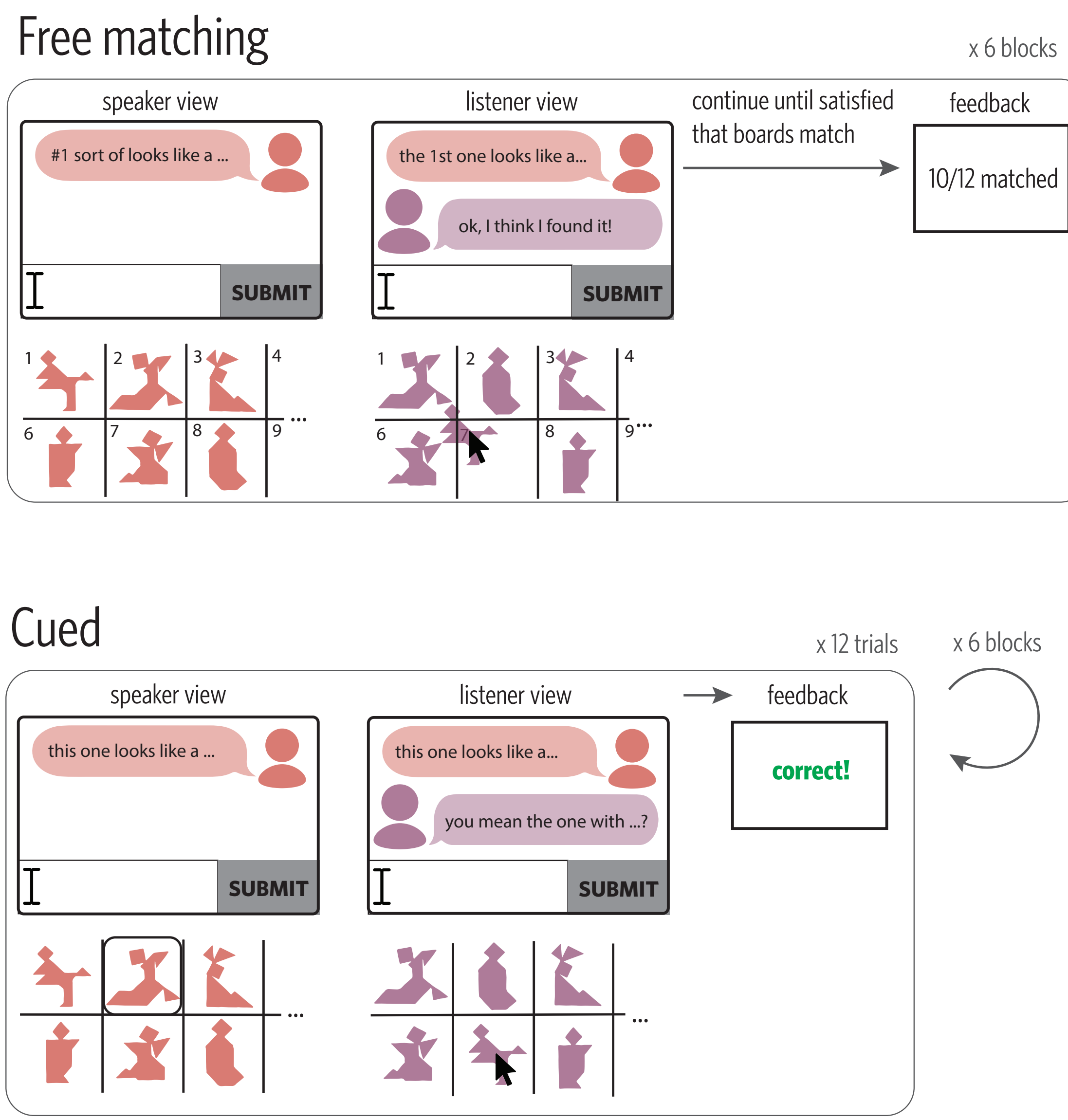

Figure 4.1: 'Free matching' and 'cued' variants of the tangrams task.

Participants

A total of 480 participants (218 in the *free-matching* version and 262 in the *cued* version) were recruited from Amazon's Mechanical Turk and paired into dyads to play a real-time communication game using the framework in (Hawkins, 2015).

Exclusion criteria

After excluding games that terminated before the completion of the experiment due to server error or network disconnection (40 in *free matching* and 33 in *cued*), as well as games where participants reported a native language different from English (2 in *free matching* and 3 in *cued*), we implemented an additional exclusion criterion based on accuracy. We used a 66/66 rule, excluding pairs that got fewer than 66% of the tangrams correct ($\geq$ 8 of 12) on more than 66% of blocks ($\geq$ 4 of 6). While the most pairs were near ceiling accuracy by the final round, this rule excluded 11 in *free matching* and 8 in *cued* who appeared to be guessing or rushing to completion. After all exclusions, we were left with a *free matching* corpus containing a total of 8,639 messages over 56 complete games and a *cued* corpus containing 9,164 messages over 83 games.

Stimuli & Procedure

On every trial, participants were shown a 6×2 grid containing twelve tangram shapes, reproduced from (H. H. Clark & Wilkes-Gibbs, 1986). After passing a short quiz about task instructions, participants were randomly assigned the role of either 'director' or 'matcher' and automatically paired into virtual rooms containing a chat box and the grid of stimuli. Both participants could freely use the chat box to communicate at any time.

In the *free-matching* version, our procedure closely followed (H. H. Clark & Wilkes-Gibbs, 1986). The director and matcher began each round with scrambled boards. The director's tangrams were fixed in place, but the matcher's could be clicked and dragged into new positions. The players was instructed to communicate through the chat box such that the matcher could rearrange their shapes to match the order of the director's board. When the players were satisfied that their boards matched, the matcher clicked a 'submit' button that gave players batched feedback on their score (out of 12) and scrambled the tangrams for the next round. After six rounds, players were redirected to a short exit survey. Cells were labeled with fixed numbers from one to twelve in order to help participants easily refer to locations in the grid (see Fig. 4.1).

While this replicated design allows for highly naturalistic interaction, it poses several problems for text-based analyses. First, utterances must contain not only descriptions of the tangrams but also information about the intended location (e.g. '*number 10* is the ...'). Additionally, because there were no constraints on the sequence, participants can revisit tangrams out of order or mention multiple tangrams in a single message, making it difficult to isolate exactly which utterances referred to which tangrams without extensive hand-annotation. Finally, the design of the 'submit' button made it easy for players to occasionally advance to the next round without referring to all 12 tangrams.

For the *cued* version, then, we designed a more straightforwardly sequential variation on the task where speakers are privately cued to refer to targets one-by-one and feedback is given on each round (see Fig. 4.1); this allows us to straightforwardly conduct analyses at the tangram-by-tangram level. On each trial, one of the twelve tangrams was privately highlighted for the director as the *target*. Instead of clicking and dragging into place, matchers simply clicked the one they believed was the target. They were not allowed to click until after a message was sent by the speaker. We constructed a sequence of six blocks of twelve trials (for a total of 72 trials), where each tangram appeared once per block. Because targets were cued one at a time, numbers labeling each square in the grid were irrelevant and we removed them. The context of tangrams was scrambled on every trial, and participants were given full, immediate feedback: the director saw which tangram their partner clicked, and the matcher saw the intended tangram.

Data pre-processing

We used a three step pre-processing pipeline to prepare our corpus for subsequent analyses. Unless otherwise noted, we used the open-source Python package `spaCy` to implement all NLP tasks.

1. Spell-checking and regularization: We conservatively extracted all tokens that did not exist in the vocabulary of the smallest available ($\sim$ 50,000 word) `spaCy` model and passed them through the SymSpell spell-checker [‡]. These suggested corrections were then sequentially presented to the first author and either accepted or overridden at their judgement. This process constructed a reproducible spell-correction dictionary we applied to our dataset.

2. Cleaning unrelated discourse: Because we allowed our participants to interact in real-time through the chat box, many pairs produced text unrelated to the task of referring to the current target (e.g. greeting one another, asking personal questions, commenting on the length of the task or the results of previous rounds). We wanted to ensure that our structural results were not confounded by patterns in this kind of discourse across the task, and that the semantic content we observe on a particular trial is in fact being used to refer to the current target rather than task-irrelevant topics or, as we found in some cases, referring to other tangrams while debriefing previous errors. We therefore applied a manual pass applying a rubric that any text not directly referring to the current target is removed. For example, utterances like "this is the one we got wrong last time" were kept in because they were referring to a property of the current tangram, but utterances like "good job" and "they'll go quicker if you remember what I say!" are not. This process also created a reproducible JSON.

3. Collapsing multiple messages within a round: Finally, some speakers used our chat box like an texting interface, hitting the enter key between every micro-phrase of text. This made it difficult to interpret the output of syntactic parses. We therefore collapsed repeated messages by a participant within a round into a single message by inserting commas between successive messages. We chose to use commas because it tends to maintain grammaticality and does not inflate word counts.

[‡] `https://github.com/wolfgarbe/SymSpell`

4.2 Results: characterizing the dynamics of content

The inferential account laid out in earlier chapters makes three key predictions about how speakers change the content of their referring expressions over time. *First*, if participants are influenced by pragmatic pressures to be informative, the labels that conventionalize should not be a random draw from the initial description. Instead, we predict that more *distinctive* words in initially successful labels (e.g. words used exclusively to describe one tangram) will be more likely to remain in later descriptions. *Second*, due to sources of variability in the population of speakers, we predict that the referring expressions used by different pairs will increasingly diverge to different, idiosyncratic labels. In other words, different pairs will find different but equally successful equilibria in the space of possible linguistic conventions. *Third*, as speakers learn and gradually strengthen their expectations about how their partner will interpret their referring expressions, the labels used within each pair for each tangram with stabilize. In other words, once there is evidence that a particular label is successfully understood, there is little reason to deviate from it. Because these analyses depend on tangram-level resolution, we only examine the "cued" dataset in this section.

4.2.1 Initially distinctive words are more likely to conventionalize

We begin by investigating *which* content is dropped and which is preserved. Which computational principles may allow us to predict whether a particular word in a speaker's initial description of an object will become established as a convention for referring to it on later rounds? In previous chapters, we discussed two principles that are particular relevant for this question. First, if speakers are attempting to be informative in a particular context of other tangrams then the Gricean maxim of quality suggests that a good referring expression is one that applies more strongly to the target than to the distractors. Properties that are shared in common across multiple objects are poor candidates for conventions that must distinguish among them. Second, the principles of cross-situational meaning adaptation suggest that these informativity considerations will be strengthened through learning. The exclusive usage of a word with one tangram and no others should reinforce the specificity of that meaning in the local discourse context, even if the listener may be *a priori* willing to extend it to other targets. Conversely, if a particular word has been successfully used with several different referents, its specificity may be weakened in the local context. Putting these principles together, we hypothesized that more *initially distinctive* words would be more likely to conventionalize.

For each pair of participants, we quantified the distinctiveness of a word w as n_w: the number of tangrams that it was used to describe on the first repetition. A word that is only used in the description of a single tangram (e.g. a descriptive noun like "rabbit") would be very distinctive, while a word used with all 12 tangrams (e.g. an article like "the") would be not distinctive at all. While this formulation is the most transparent to state in words, it is equivalent (up to a constant) to two popular and theoretically motivated measures of distinctiveness used in natural language processing (Salton & Buckley, 1988). The first is *term frequency-inverse document frequency* (tf-idf; Sparck Jones, 1972), which multiplies the term frequency $tf(w, d)$ of a word w in a document d by a "global" term $\log(N/n_w)$ where N is the total number of documents and n_w is the number of documents containing w. In our case, the "documents" are just the referring expressions used for a distinct tangram on the first round, so $N = 12$ and we can take $tf(w, d)$ to be a boolean for simplicity: 1 if the word occurs, 0 if it does not. We can thus retrieve our simpler measure by exponentiating, dividing by 12, and taking the inverse. The second is *positive point-wise mutual information* (PPMI). Point-wise mutual information compares the joint probability of a word occurring with a particular tangram to the probability of the two occurring independently:

$$PMI_{word,tangram} = \log \frac{P(word, tangram)}{P(word)P(tangram)}$$

Positive point-wise mutual information is given by $\min(0, PMI)$, restricting the lower bound to 0.

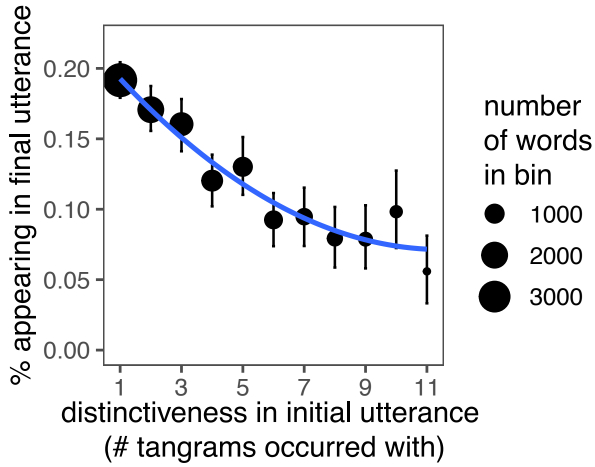

Figure 4.2: More distinctive words are more likely to conventionalize. Points represent estimates of the mean probability of conventionalizing across all words with a given distinctiveness value. Size of points represent the number of words at that value. Curve shows regression fit; error bars are bootstrapped 95% CIs.

It can be shown for our case that *tf-idf* is the maximum likelihood estimator for PPMI: the numerator reduces to a boolean when we only have one observation per tangram (Robertson, 2004).

Given this simple but principled measure of word distinctiveness at the speaker-by-speaker level – the number of tangrams it was initially used with – we were interested in the extent to which it accounts for conventionalization, the probability that a word in the speaker's initial description is preserved until the end of the game. More than half of the words used to refer to a tangram on the final round (57%) appeared in the initial utterance. [§] We thus restricted our attention to this subset of words, coding them with a 1 if they later appeared in the final round and 0 if they did not. We then ran a mixed-effects logistic regression including a fixed effect of initial distinctiveness (using the transformed *tf-idf* for stability) and maximal random effect structure with intercepts and slopes for each tangram and pair of participants. We found a significant positive effect of distinctiveness:

[§]The 43% of final round words that did not exactly match were often synonyms or otherwise semantically related to words used on the first round, e.g. "foot" on the first round vs. "leg" on the last. In other cases, the labels used at the end were introduced after the first repetition, e.g. one pair only started using the conventionalized label "portrait" on repetition 3.

words that were used with a larger number of tangrams on the first round were less likely to conventionalize, $b = -0.23$, $z = -6.1$ (see Fig. 4.2). Similar results are found using the derived measure of *tf-idf*.

Finally, we conducted a non-parametric permutation test. For each speaker and tangram, we *randomly sampled* a word from the initial utterance and computed the mean probability of this word also being used on the final round. Repeating this procedure 1000 times yielded a null distribution ranging from 2.5% to 6.6%. However, if we instead sample from the words with *maximal distinctiveness*, we obtained a distribution ranging from 24% to 31%, which is non-overlapping with the null distribution. Thus, if we must make a bet on which words will become conventionalized, placing our bet on the most distinctive ones will yield much higher returns.

4.2.2 Conventions diverge across pairs and stabilize within pairs

To jointly examine our other predictions about the dynamics of content, we introduce two different quantitative measures of similarity: one based on properties of the discrete word count distribution and the other based on distances computed between continuous vector embeddings of referring expressions. Because these analyses depend on the identity of word tokens, we applied a lemmatizer to further standardize the input. Lemmatization maps multiple morphological variants (e.g. 'played,' 'playing,' 'plays') to the same stem ('play'). We do not want an observed difference between two pairs to be driven simply by different forms of the same word.

Measuring convergence and divergence with discrete word distributions We begin by examining the discrete *distribution of words* that each pair uses to refer to each tangram, excluding standard stop words. If a pair of participants converges on stable labels for a tangram, then this stability should manifest in a highly structured distribution over words throughout the game for that pair. If different speakers discover diverging conventions, this idiosyncracy should also man-

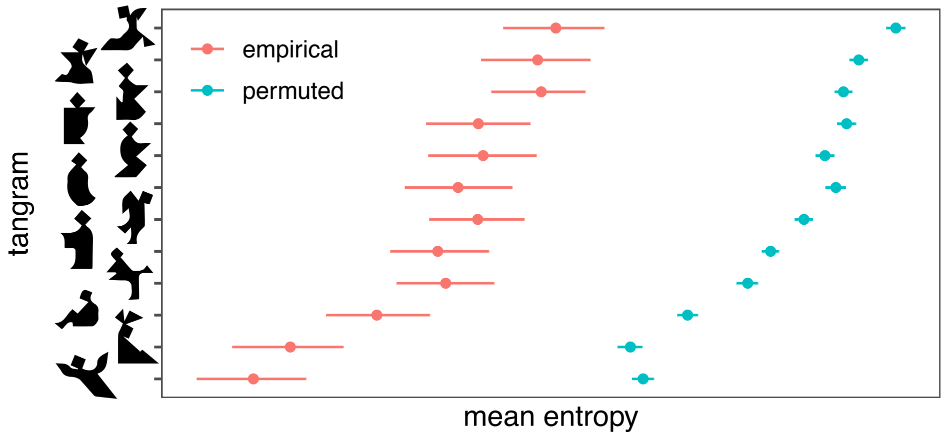

Figure 4.3: Permuting utterances across pairs increases entropy of word distribution, consistent with internal stability and multiple equilibria. Mean empirical entropy (red) and mean permuted entropy (blue) are shown for each tangram. Error bars are 95% CIs for bootstrapped empirical entropy and the permuted distribution, respectively.

ifest in differing word distributions. We formalize these intuitions by examining the information-theoretic measure of entropy:

$$H(W) = \sum_w P(w) \log P(w)$$

The entropy of the word distribution for a pair is maximized when all words are used equally often and declines as the distribution becomes more structured, i.e. when the probability mass is more concentrated on a subset of words.

To compare word distributions across games, we use a permutation test methodology. By scrambling referring expressions for each tangram across games and recomputing the entropy of the scrambled word distribution, we effectively disrupt any structure within each pair. There are two important inferences we can draw from this test. First, in a null scenario where different pairs did *not* diverge as predicted and instead every pair coordinated on roughly the same (optimal) convention for each tangram, this permutation operation would have no effect since it would be mixing together copies of the same distribution. Second, in another null scenario where pairs did not converge and instead varied wildly in the words they used from round to round, then permuting across games would also have no effect since it would simply mix together word distributions that already have high entropy. Hence, scrambling should *increase* the average game's entropy only in the case where both predictions hold: each game's idiosyncratic but concentrated distribution of words would be mixed together to form more heterogeneous and therefore high-entropy distributions.

Following this logic, we computed the average within-game entropy for 1000 different permutations of speaker utterances. We permuted utterances within rounds rather than across the entire data set to control for the fact that earlier rounds may generically differ from later rounds. Because we are permuting and measuring entropy at the tangram-level, this yields 12 permuted distributions (see Fig. 4.3). We found that the mean empirical entropy lay well outside the null distribution for all twelve tangrams, $p < .001$, consistent with our predictions of internal stability within pairs and multiple equilibria across pairs.

Measuring similarity using vector space embeddings A more direct way to quantify convergence within and divergence across different pairs is to use a continuous similarity between vector space embeddings of utterances. Although the idea of using dense vector space representations of words to measure similarity is an old one (Osgood, 1952; Landauer & Dumais, 1997; Bengio, Ducharme, Vincent, & Jauvin, 2003), recent breakthroughs in machine learning have yielded rapid improvements in these representations (e.g Mikolov, Sutskever, Chen, Corrado, & Dean, 2013; Pennington, Socher, & Manning, 2014). To quantify the dynamics of semantic context in referring expressions across and within games, we extracted the 300-dimensional GloVe vector for each word. We then averaged these word vectors to obtain a single sentence vector for each referring expression[¶]. To avoid artifacts from function words, we only included open-class content words (nouns,

[¶]Variations on such naive averaging methods are surprisingly strong baselines for sentence representations (Arora, Liang, & Ma, 2017), performing better than supervised LSTM representations or unsupervised skip-thought vectors (Kiros et al., 2015)

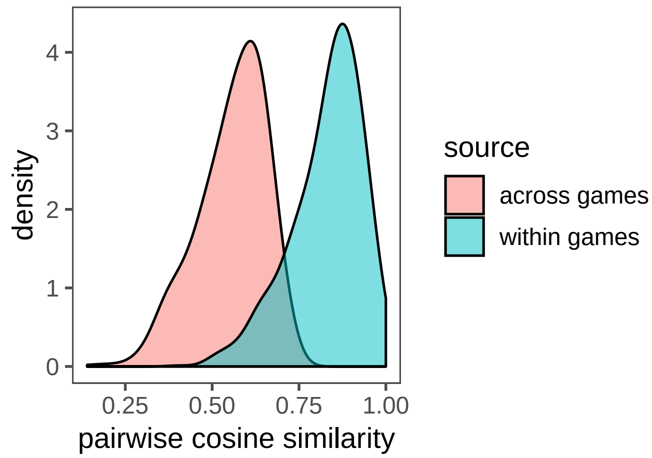

Figure 4.4: Distribution of similarities between different utterances within and across different games.

adjectives, verbs) in this average. We can then define a similarity metric between any pair of vectors $\langle u_i, u_j \rangle$. We find that our results are robust to several choices of metric, but for simplicity we will use cosine similarity

$$\cos\theta_{ij} = \frac{u_i \cdot u_j}{\|u_i\|\|u_j\|}$$

throughout the presentation below.

Utterances are more similar overall within games than between games Before examining the dynamics of these vectors, we first test the basic prediction that utterances *within* a game are more similar overall than utterances *across* games, reflecting systematic variability in how different pairs solve the referential challenge posed by the reference game. For each tangram, we computed the pairwise similarity between all utterances *within* a game and also *across* games. The distributions of these values are shown in Fig. 4.4. We estimated the distance between these distributions using the standard normalized sensitivity $d' = \frac{\mu_A - \mu_W}{\sqrt{1/2(\sigma_A^2 + \sigma_W^2)}} = 2.71$. To compare this estimated difference against the null hypotheses that within- and across-game similarities are drawn from the same distribution, we conducted a permutation test by scrambling 'within' and 'across' labels for each similarity and re-computing d' 1000 times. We found that our observed value was extremely unlikely under this null distribution, $95\% \; CI : [-0.09, 0.09], p < 0.001$.

In other words, utterances from a single pair tend to cluster together in semantic space while different pairs are more spread out in different parts of the space. This observation is consistent with our hypothesis that different pairs discover different conventions while a single pair tends to keep using a convention once established. Having established this separation between similarity distributions in aggregate, we proceed to ask more fine-grained questions about the *dynamics* through space: how do individual pairs evolve in their content over successive rounds? To more rigorously test our predictions about gradual divergence to multiple equilibria and convergence to internally stable conventions, we conducted three analyses directly on the semantic vectors.

Increasing dissimilarity from initial utterance First, we hypothesized that there was cumulative change in the semantic content of a particular pair's utterances across repetitions. Concretely, we predicted that within a particular pair of participants, utterances on later repetitions would become increasingly dissimilar from the initial utterance. We tested this prediction in a mixed-effects regression model including (orthogonalized) linear and quadratic fixed effects of the 'lag' from the first repetition (i.e. 1 for the second repetition, 2 for the third repetition, etc) as well as maximal random effects for each tangram and pair of participants. We found a significant linear decrease in similarity to the initial round as the lag becomes larger, $b = -3.5, \; t = 12.2$, as well as a significant quadratic term, $b = 1.1, t = 5.3$, suggesting that this decrease in similarity slows down over time (see Fig. 6.5A).

However, since the entire distribution of utterances may have drifted to a different region of the semantic space for generic reasons (e.g., because they were shorter overall), we compared the estimated drift *within* pairs of participants to a permuted baseline. For each target tangram, we scrambled utterances across different pairs of participants and re-ran our mixed-effects model to

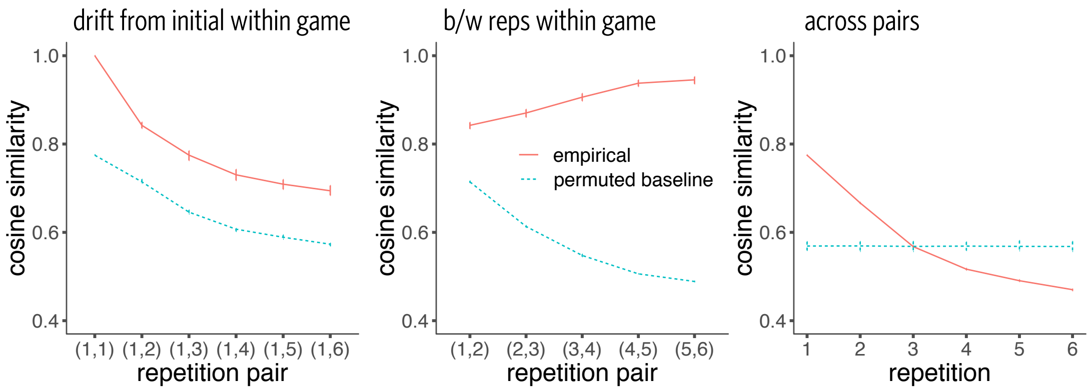

Figure 4.5: Utterances within a pair (A) become more dissimilar from initial utterance and (B) become more similar to successive utterances on later repetitions, but (C) utterances across pairs become steadily more dissimilar. Error bars are bootstrapped 95% CIs; dotted line represents permuted baseline.

obtain a null distribution representing the decrease in semantic distance from a *random* speaker's utterance on the first round. We found that this permuted baseline also showed a linear decrease over time, but our true estimate ($b = -3.5$) fell narrowly outside the null distribution of effects ($95\%\ CI = [-3.34, -3.04]$), showing that utterances by a particular speaker drifted from their own initial utterance to a slightly greater degree than would be expected due to generic differences between utterances made at different timepoints in an interaction [‖]. This difference is likely a consequence of random utterances from different speakers being more dissimilar even on *early* repetitions, thus depressing the overall slope.

Increasing internal consistency within interaction As speakers modified their utterances across successive repetitions, we additionally hypothesized that they would converge on increasingly consistent ways of referring to each tangram. To test this prediction, we computed the semantic similarity between successive utterances produced by each speaker when referring to

[‖]Both here and for the permuted baselines in the subsequent two analyses we needed to simplify the random effects structure to contain only random intercepts due to convergence issues over the large number of permutations. However, we were interested in the coefficient estimate rather than statistical significance in these permuted models, and estimates appeared stable across different random effects structures.

same tangram (i.e. repetition k to $k+1$). A mixed-effects model with linear and quadratic fixed effects of repetition number and maximal random effects for both tangram and pair of participants showed that similarity between successive utterances increased substantially throughout an interaction ($b = 2.7,\ t = 10.9$; Fig. 6.5B). The quadratic term was not significant ($b = -0.4,\ t = -1.8$). Again, we compared our empirical estimate of the magnitude of this trend to a null distribution of slopes estimated by scrambling utterances across pairs and re-running the regression model.The estimated slope fell outside this null distribution, for which similarity was strongly *decreasing*, $CI = [-5.9, -5.4]$, providing evidence that increasingly consistent ways of referring to each object manifested only for series of utterances produced within the same interaction.

Increasingly different content across interactions Finally, we predicted that the way different pairs refer to the same tangram would become increasingly dissimilar from each other across repetitions, gradually diverging into different equilibria. We tested this prediction by computing the mean pairwise similarity between utterances used by different speakers to refer to the same object. The large sample of pairwise similarities ($N = 257,040 = 12$ tangrams $\times$ 6 repetitions $\times$ $\frac{85 \cdot 84}{2}$ distinct pairs) presented both advantages and disadvantages. On one hand, we could obtain highly reliable estimates of mean similarity. On the other hand, larger random-effects structures led to convergence problems. We therefore ran a mixed-effects regression model including linear and quadratic fixed effects of repetition number including maximal random effects only at the tangram-level. We found a strong negative linear fixed effect of repetition on between-game semantic similarity ($b = -50.7, t = 16.8$) as well as a significant quadratic effect ($b = 16.1, t = 12$), indicating that this divergence slows over time as each pair stabilizes, (see Fig. 6.5C). We again conducted a permutation test to compare this t value with what would be expected from scrambling utterances across repetitions for each pair and target. We found that the estimated slope was highly unlikely under this distribution ($CI = -2.5, 2.9],\ p < 0.001$).

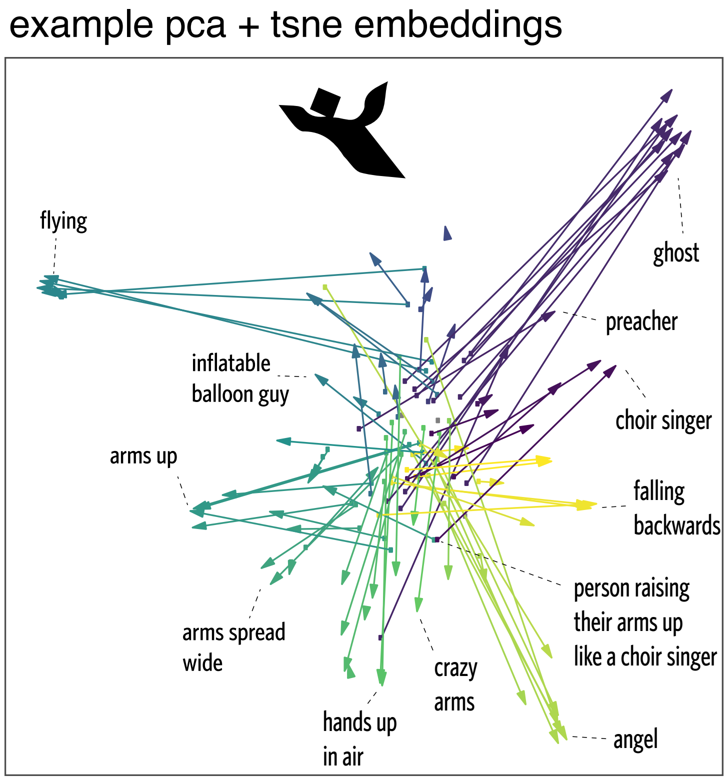

Figure 4.6: 2D projection of semantic embeddings for example tangram. Each arrow represents the trajectory between the first round to last round for a distinct pair of participants. Color represents the rotational angle of the final location to more easily see where each pair began. Annotations are provided for select utterances, representing different equilibria found by different participants.

Visualizing trajectories through vector space Finally, to better understand the changes uncovered by these analyses of utterance embeddings, we visualize the trajectories taken by each pair of participants when referring to a particular example tangram, annotating utterances in several parts of the space. First, we took the first 50 components recovered by running Principal Components Analysis (PCA) on the 300-dimensional utterance embeddings. We then use t-SNE (Maaten & Hinton, 2008) to stochastically embed the lower-dimensional PCA representation of each utterance in a common 2D vector space. In Fig. 4.6, each arrow connects the first and last utterance a particular pair used to refer to this tangram.

We observe that the initial utterances of each game tend to cluster tightly near the center of the space and the final utterances are *dispersed* more widely around the edges. This pattern is consistent with the hypothesis that different speakers overlap more in the content of their early descriptions before honing in on more distinctive different equilibria later in the game (see 4.2.1). For this particular tangram, there were a handful of semantically distinct labels that served as equilibria for multiple pairs ("ghost," "flying," "angel") as well as many more idiosyncratic labels. Pairs often initially mentioned multiple properties (e.g. "person raising their arms up like a choir singer") before breaking the symmetry and collapsing to one of these properties ("choir singer").

4.3 Results: characterizing the dynamics of structure

In the previous section, we examined the dynamics of semantic content. We found that pairs converged systematically on distinctive words but different pairs discovered different solutions to the same coordination problem. Here, by contrast, we examine the dynamics of *structure*, testing the extent to which different pairs share more abstract commonalities in the structural changes of their referring expressions over time, even while differing in their content. In particular, we examine *how* different pairs reduce the length of their utterances. What sequence of transformations is applied to reduce long initial descriptions into shorter final ones?

4.3.1 Dialogue between speaker and listener

Before focusing our analysis on the dynamics of how *directors* transform their referring expressions over time, we first on the broader structure of bi-directional dialogue exchanges. Conceptual pacts are formed *collaboratively* (H. H. Clark & Wilkes-Gibbs, 1986): directors and matchers engage in a bi-directional process where matchers ask follow-up questions, suggest corrections, and acknowledge or verbally confirm their understanding through a backchannel. In the absence of feedback,

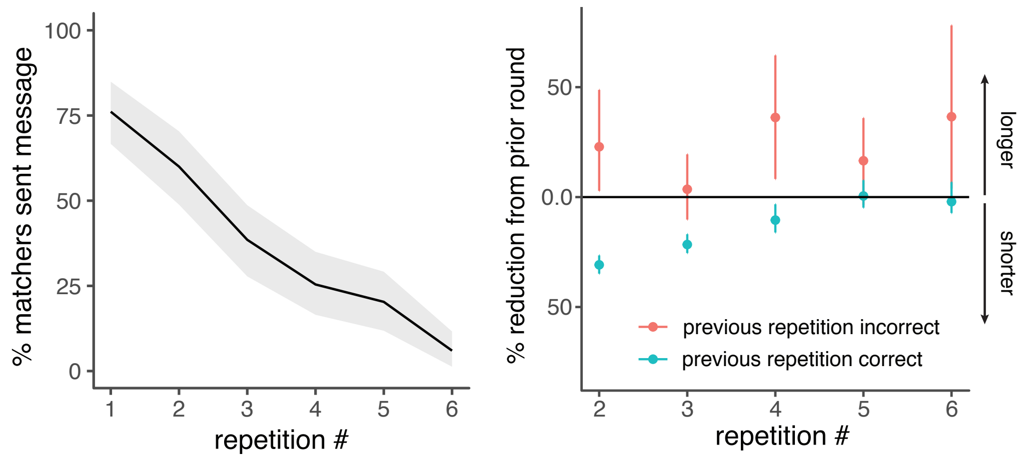

Figure 4.7: (A) Listeners reply with interactive feedback less often over repetitions, and (B) speakers are sensitive to listener response feedback, increasing message length on the subsequent repetition of a tangram after an error is made.

descriptions may not necessarily get shorter (Krauss & Weinheimer, 1966; Garrod, Fay, Lee, Oberlander, & MacLeod, 2007).

While we automatically supplied minimal feedback about the listener's response each round, we predict that the additional feedback from listener backchannel replies should be highest on the first round and drop off once meanings are agreed upon. To test this prediction, we coded whether the listener sent a message or not on each trial and fit a mixed-effects logistic regression model with a fixed effect of repetition, random intercepts and slopes for each pair of participants, and a random intercept for each target. We found that the probability of the listener sending a message decreased significantly over the game ($b = -0.84, t = -9.1, p < 0.001; Fig. 4.7A$). In aggregate, 76% of listeners send at least one message in the first repetition block, but only 6% sent a message in the last block. These rates found in our online text-based replication are lower overall than in-person lab experiments, but we nonetheless strongly replicated the overall trend.

Next, we examined the extent to which speakers were sensitive to listener response feedback in modulating their utterances. If the listener failed to select the correct target, the speaker may take this as evidence that their description was insufficient and attempt to provide more detail the next time they must refer to the same tangram. If the listener is correct, on the other hand, the speaker may take this as evidence of understanding and reduce their level of detail on future repetitions. We tested these predictions by comparing the proportional change in utterance length on the repetition after an error against the change in length after a correct response (i.e. $(n_t - n_{t-1})/n_{t-1}$). This measure could be positive, indicating a net increase in utterance length, or negative, indicating a reduction.

We fit a mixed-effects regression model predicting this measure with an effect-coded categorical fixed effect of previous round feedback and a (centered) continuous effect of repetition number, including random intercepts and effects of feedback for each speaker. We found a significant main effect of feedback, even controlling for repetition: utterance length changed more in the shorter direction after correct responses than after negative responses, $b = -0.18, t = -6.2$ (see Fig. 4.7B). Indeed, speakers were more likely on average to *add* words on the repetition after an error at any point in the game. Because repetitions of the same tangram were spaced out and errors were relatively rare, this effect is unlikely to simply reflect heightened attention on trials after an error. Instead, this pattern of results is consistent with sensitivity to tangram-specific evidence of the listener's understanding.

4.3.2 Understanding reduction

Next, we turn to a set of analyses examining the reduction in utterance length over the course of the experiment. At the coarsest level, we find that the mean number of words used by speakers decreases over time (see Fig. 4.8) in both the free matching and cued variants of the task. Free matching required more words overall because participants needed to additionally mention which tangram they were referring to (i.e. "number 3 is the ..."). This result replicates the highly reliable reduction effect found throughout the literature on repeated reference games (e.g Krauss & Weinheimer, 1964; Bren-

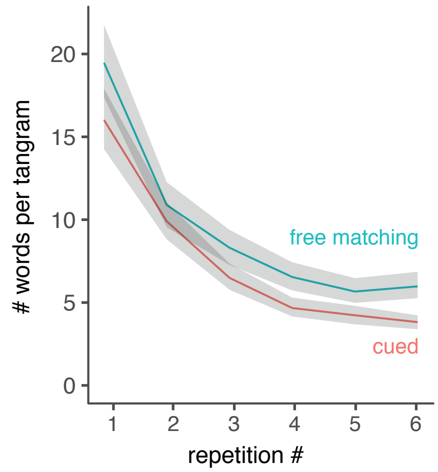

Figure 4.8: Similar reduction in # words per tangram for both variants of the task. Error ribbons are 95% confidence intervals.

nan & Clark, 1996), including our own results using COCO images in Chapter 3. Perhaps because of the text-based (vs. spoken) interface, participants in our task used fewer words overall than reported by (H. H. Clark & Wilkes-Gibbs, 1986). The following analyses break down this broad reduction into a finer-grained set of phenomena on the *cued* corpus.

Reduction in parts of speech The first level of granularity concerns which *kinds* of words are most likely to be dropped. We used the SpaCy part-of-speech tagger (Honnibal & Montani, 2019) to count the number of words belonging to different parts of speech in each message. In Fig. 4.9A, we show the shifting proportions of different parts of speech at each repetition, We find that nouns account for proportionally more of the words being used over time, while determiners and prepositions account for fewer. To test which kinds of words are more likely to be dropped, we measured the percent reduction in the number of words in each part of speech from the first round to the sixth round. We find that determiners ('the', 'a', 'an') are the most likely class of words to be dropped (90%) and nouns ('dancer', 'rabbit') are the least likely to be dropped (61%).More gener-

	unigrams	bigrams	trigrams
#1	a	look like	look like a
#2	the	like a	look like -PRON-
#3	-PRON-	to the	to the right
#4	like	-PRON- be	like -PRON- be
#5	be	this one	like a person
#6	look	the right	to the left
#7	on	the left	this one look
#8	one	like -PRON-	one look like
#9	with	on the	this one be
#10	to	with a	-PRON- look like
#11	and	a person	look like someone
#12	right	on top	diamond on top
#13	this	in the	in the air
#14	of	a diamond	on top of
#15	head	have a	a diamond on

Table 4.1: Top 15 unigrams, bigrams, and trigrams with the highest numeric reduction from first round to last round. Text lemmatized before n-grams computed.

ally, closed-class parts of speech, including function words like determiners, are strictly more likely to be dropped than open-class parts of speech (Fig. 4.9B). Because open-class parts of speech are statistically more likely to supply *distinctive* words than closed-class parts of speech, these structural considerations may contribute to the patterns in distinctiveness reported in section 4.2.1.

One possible interpretation of these findings is that reduction may be driven mostly by the loss of function words as speakers shift to a less grammatical shorthand over the course of the task. However, when examining the n-grams most likely to be dropped (see Table 4.1), we find that many of the most dropped closed-class words are used to form conjunctions ('and') or prepositional phases ('of', 'with'). Others are modifiers ('the right ...'). These examples suggest an alternative explanation: the higher reduction of closed-class function words may be a consequence of entire meaningful grammatical units being dropped at once. If initial descriptions tend to be syntactically complex, combining multiple partially redundant sources of information for identifying the target, then the

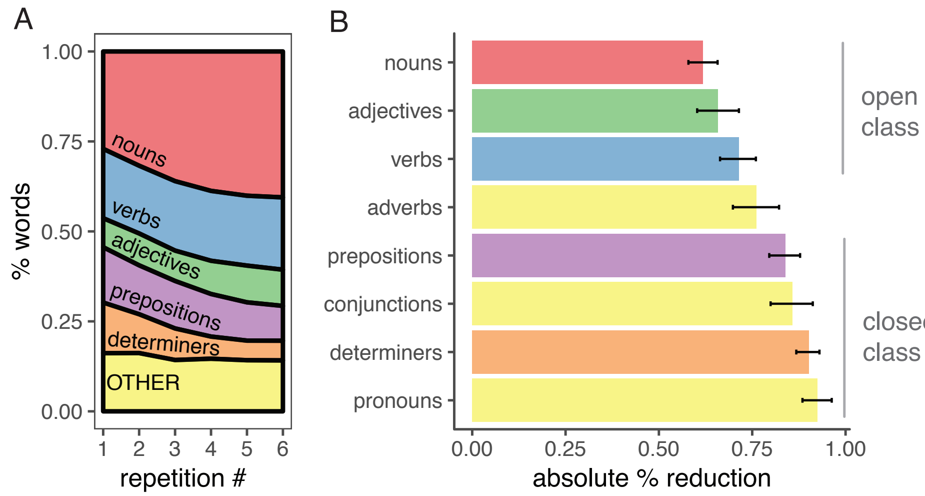

Figure 4.9: (A) proportion of words in different part of speech on each repetition. (B) Closed-class parts of speech are more likely to be dropped than open-class parts of speech. Error bars are bootstrapped 95% confidence intervals.

speaker may omit entire modifying clauses with additional evidence of a listener's understanding.

Reduction in syntactic constituents We explicitly examined this hypothesis by examining whether dropped words tend to come from the same syntactic units, relative to a random deletion baseline. We quantified the extend to which dropped words 'cluster' by examining dependency lengths between the dropped words (Jurafsky & Martin, 2014; Futrell, Mahowald, & Gibson, 2015). Specifically, we compared each referring expression to the one produced on the subsequent repetition to determine which words were dropped. Then we looked up each pair of dropped words and found the shortest path between them in the dependency parse tree (see Fig. 4.10). We then took the mean dependency length across all such pairs in all games.

This empirical 'syntactic clustering' statistic was then compared to a random baseline. For this baseline, instead of examining dependency lengths for the words that were actually dropped, we randomly sampled the same number of words from the referring expression and computed the dependency length between them. We repeated this procedure 100 times to obtain a null distribution

Figure 4.10: Example dependency parse for referring expression. If the sub-phase "arms out in front" were dropped, we would find a mean dependency length of 1.33 among the dropped words.

of the mean dependency length that would be expected if words were being dropped *randomly* from anywhere in the message.

We found a mean empirical dependency length of 4.22, which lay outside the null distribution ($95\% CI : [4.38, 4.44]$), indicating a small but reliable effect of syntactic clustering. The words that were dropped tended to be closer to one another in the dependency parse than expected by chance. Other statistics such as the minimum dependency path or the raw distance in the sequence of words gave similar results. This result accords with early observations by (Carroll, 1980), which found that in three-quarters of transcripts from (Krauss & Weinheimer, 1964) the short names that participants converged upon were prominent in some syntactic construction at the beginning, often as a head noun that was initially modified or qualified by other information.

4.4 Discussion

In this chapter, we characterized the dynamics of convention formation in a classic repeated reference game paradigm. First, we found that pairs of participants systematically tend to conventionalize words that are more distinctive in the initial context. Using both discrete and continuous measures of semantic content, we found this process leads to stable usage within pairs but multiple equilibria across different pairs. Second, speakers reduce the length of their descriptions over the course of the task in a way that is sensitive to evidence of listener understanding and syntactically structured to omit entire meaningful units. In sum, these results establish new benchmark phenomena that computational models of adaptation and convention formation must account for.

Discrete vs. continuous semantic measures Because our results hinged on our quantification of semantic content, it is worth noting some advantages and disadvantages of discrete measures compared to continuous vector space measures. A key advantage of measures based on the word distribution is that the entropy is not dependent on any particular choice of pre-trained vector embedding. Due to biases in the training corpora, vector representations also may not capture some of the more idiosyncratic conventions that participants converge on (e.g. "YMCA" or "zig zag" or "Frank" – short for "Frankenstein"). To the extent we find converging results, the discrete measure may address concerns about the quality of the continuous representation.

A key disadvantage is that the entropy is sensitive to the support of the word distribution — the vocabulary present in the corpus on a particular round — and thus does not have a natural scale. While directly measuring divergence between word distributions at different repetitions and between different pairs is technically possible, their sparsity makes this approach not as informative at these finer granularities of analysis. Many pairs use entirely disjoint sets of words, and on later rounds, the distribution may only contain one or two words. Further, because it is based entirely on the frequency of tokens, it may treat even close synonyms as entirely distinct tokens in the word distribution. Thus, these two approaches provide complementary evidence for the dynamics of content.

Symmetry-breaking Our results in this chapter raise a subtle cognitive question about classic definitional notions of arbitrariness in convention (Lewis, 1969), which hold that there must counterfactually exist an alternative solution to the coordination problem for any particular solution to be conventional. How can such systematicity in the formation process co-exist with conventionality? The symmetry of different solutions must break somewhere to account for the empirical existence of many alternative but equally successful referential conventions at the population level, but our results are consistent with two different cognitive realities *within* games.

One possibility is that this arbitrariness is a product of substantial variability in a population of fairly rigid speakers. That is, each individual speaker may have strong but idiosyncratic initial preferences for how to refer to each tangram. They may begin with additional elaboration given their representation of uncertainty about whether these strong preferences are shared, but without strong evidence that their partner cannot understand, they will persist with their preferred and premeditated label.

A second possibility is that the population of speakers is homogenous but only weakly constrained by prior preferences. That is, speakers may not only be uncertain of which messages their partner will understand, but are themselves unclear on an appropriate way to refer to these unfamiliar objects. If speakers initially sample from a variable distribution, and then from their updated distribution on subsequent rounds, different pairs may end up in different equilibria due to the path-dependence of the process. In this way, the symmetry may be broken through randomness in the sampling step (or in the updating step, if learning is stochastic.) These possibilities are not mutually exclusive. Our experiments ruled out the possibility of a homogeneously rigid population, but it is possible that some speakers have strong preferences while others do not.

While prior empirical work has indirectly tested expectations and prior preferences – for instance, by asking speakers to either produce descriptions for others or for themselves in the future (Fussell & Krauss, 1989a) – an important direction for further work is to design experiments that disentangle these possibilities. For instance, a Bayesian truth serum approach (Prelec, 2004) could estimate both an individual's own subjective preferences and their expectations about whether these would be shared by others. More broadly, following recent attempts to estimate the true variability across speakers in phonetic properties of speech production (Kleinschmidt, 2019), it would be valuable to estimate how much variability in semantic expectations there really is in the population (Furnas et al., 1987).

5

Emerging abstractions: Lexical conventions are shaped by communicative context

Natural languages provide speakers with remarkable flexibility in the labels they may use to refer to things (Brown, 1958; Cruse, 1977). On top of an abundance of expressions made available by syntactic combination and semantic compositionality (Partee, 1995), we have a number of overlapping and nested terms in our lexicon. *Fido*, *Dalmatian*, *dog*, and *animal* can all reasonably be used to talk about the same entity at different levels of abstraction. How these overlapping meanings are learned, and why speakers choose different levels of specificity in different contexts, is increasingly well-understood (e.g. Xu & Tenenbaum, 2007; Graf et al., 2016) but there remains a more fundamental question about the structure of our conventions. Why and how do different levels of abstraction become conventionalized in the first place?

In Chapter 4, we found in an observational corpus that more initially distinctive words were more likely to conventionalize. Because distinctiveness depends on the context of other objects, we hypothesized that this context shaped the conventionalization process through the mechanism of pragmatic reasoning. While our hypothesis about ad hoc conventionalization specifically concerned the shorter timescales of dyadic interaction, similar pressures may operate over the multi-generational timescales of cultural evolution. Better understanding the effect of context on local conventions rapidly formed by adaptive agents over extended interactions may therefore be valuable for understanding how *languages* are globally shaped by communicative constraints.

Recent computational approaches to language evolution have argued that the lexical conventions of languages balance simplicity, or learnability, with the communicative needs of their users over longer timescales. A key prediction is that the lexicon of a *group* of language users should be sensitive to the pragmatic demands of their environment. This optimal expressivity hypothesis accounts well for the lexical distributions found in natural languages across semantic domains like color words and kinship categories (Regier, Kemp, & Kay, 2015; Gibson et al., 2017), as well as the compositional systems that emerge under iterated learning with communication in the lab (Winters, Kirby, & Smith, 2014; Kirby, Tamariz, Cornish, & Smith, 2015). 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, Carstensen, & Kemp, 2016).

Still, there are several limitations to the current evidence for this hypothesis. First, much of the relevant evidence is observational, aggregated at the level of overall language statistics, not by directly manipulating the contextual conditions of individual language users. Second, previous experimental studies have largely focused on the functional outcomes of an iterated learning process, but have

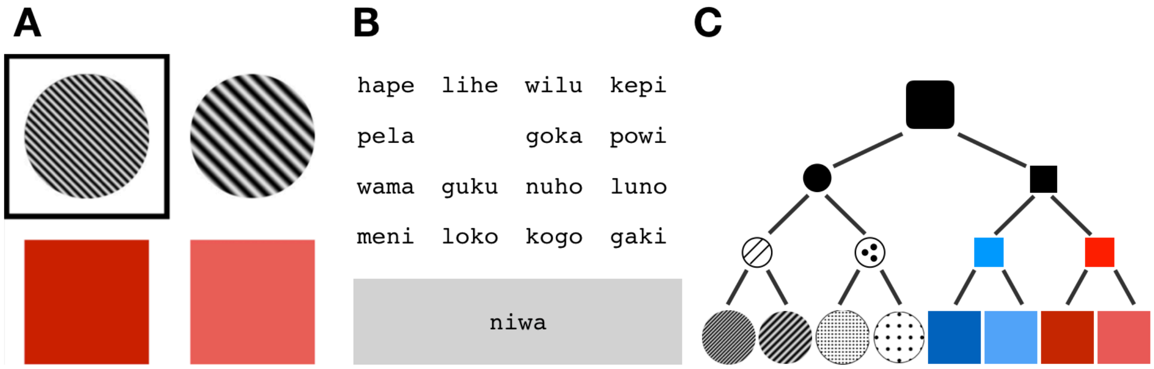

Figure 5.1: (A) Example of *fine* context where one of the distractors belongs to the same fine-grained branch of the hierarchy as the target (i.e. another striped circle), so any abstract label would be insufficient to disambiguate them. The target is highlighted for the speaker with a black square. (B) Drag-and-drop chat box interface. (C) Hierarchical organization of stimuli.

not grounded the results of this process in a cognitive, mechanistic account of lexical adaptation and convention-formation among individual agents. Finally, the phenomenon of reference taxonomies poses a further theoretical challenge: why do languages have hierarchies of terms instead of flatly partitioning the space into category labels as previous work has assumed?

Under the logic of a local efficiency/informativity tradeoff, we make two predictions about the emergence of abstractions in dyads. First, we expect that communicative pressures for informativity should lead to the lexicalization of specific names when fine distinctions must be drawn. Second, abstractions should become lexicalized precisely when the relevant distinctions are at coarser levels of the conceptual hierarchy. For example, we are often called upon to make fine distinctions between people in our social circles, hence lexicalizing efficient names for each individual; when referring to green beans or paper towels, however, we can get away without such specific terms – we are rarely called upon to disambiguate between entities.

Here, we develop an experimental paradigm and analytic approach to examine the causal factors driving the emergence of lexical conventions in real-time. We manipulated context in a repeated reference game where pairs of participants interactively coordinated on an artificial language from scratch. Even though a complete communication system containing a distinct word for each object is feasible and sufficient for all contexts, we find that abstractions begin to emerge when fine-grained distinctions are not necessary.

5.1 Experiment: Repeated reference game

Participants

We recruited 278 participants from Amazon Mechanical Turk to play an interactive, multi-player game using the framework described in Hawkins (2015). Pairs were randomly assigned to one of three different conditions, yielding between $n = 36$ and $n = 53$ dyads per condition, after excluding participants who disconnected before completion.*

Procedure & Stimuli

Participants were paired over the web and placed in a shared environment containing an array of objects (Fig. 1A) and a 'chatbox' to send messages from a randomly generated vocabulary (Fig. 1B). On each of 96 trials, one player (the 'speaker') was privately shown a highlighted target object and allowed to send a single word to communicate the identity of this object to their partner (the 'listener'), who subsequently made a selection from the array. Players were given full feedback, swapped roles each trial, and both received bonus payment for each correct response.

The objects that served as referents were designed to cluster in a fixed three-level hierarchy with shape at the top-most level, color/texture at the intermediate levels, and frequency/intensity at the finest levels (see Fig. 1C). Each communicative context contained four objects. Distractors could differ from the target at various level of the hierarchy, creating different types of contexts defined by

*All materials and data are available at https://github.com/hawkrobe/conventionalizing_hierarchies; planned sample sizes, exclusion criteria, and behavioral analysis plan were pre-registered at https://osf.io/2hkjc/.

the finest distinction that had to be drawn. We focus on two: *fine* trials, where the closest distractor belongs to the same fine-grained subordinate category (e.g. another striped circle; see Fig. 1A), and *coarse* trials, where the closest distractor belongs to a coarser level of the conceptual hierarchy (e.g. dotted circle instead of striped circle).[†] Fixed arrays of 16 utterances (enough to allow the potential for full expressibility) were randomly generated for each pair (and held constant across trials) by stringing together consonant-vowel pairs into pronounceable 2-syllable words (see Fig. 1B).

Critically, we manipulated the statistics of the context in a between-subjects design to test the effect of communicative relevance on lexicalization. In the pure *fine* and *coarse* conditions, all targets appeared in fine or coarse contexts, respectively; in the *mixed* condition, the two context types were equally likelySequences of trials were constructed by randomly shuffling targets and trial types within blocks and ensuring no target appeared more than once in a row.

In addition to behavioral responses collected over the course of the game, we designed a post-test to explicitly probe players' final lexica. For all sixteen words, we asked players to select all objects that a word can refer to (if any), and for each object, we asked players to select all words that can refer to it (if any). Using a bidirectional measure allows us to check the internal validity of the lexica reported.

5.1.1 Results

Partners successfully learn to communicate

Although participants in all conditions began with no common basis for label meanings, performing near chance on the first trial (proportion correct $= 0.19$, 95% CI $= [0.13, 0.27]$), most pairs were nonetheless able to coordinate on a successful communication system over repeated interaction (see

[†]Even coarser trials with super-ordinate distractors (e.g. a circle target among three square distractors) were logically possible but would have introduced several experimental confounds; we opted to leave these trial types out of our design and conduct the minimal manipulation.

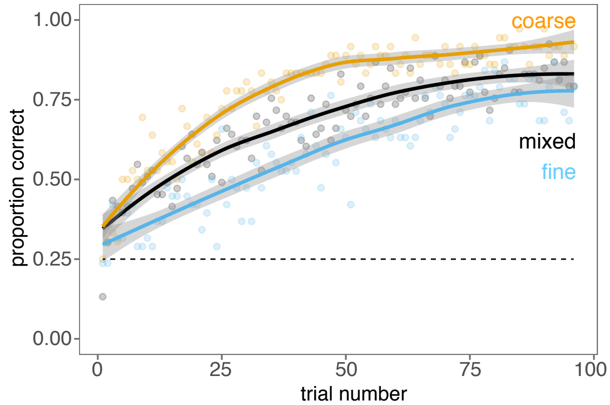

Figure 5.2: Players learn to coordinate on a successful communication system. Each point is the mean proportion of correct responses by listeners; curves are nonparametric fits.

Fig. 5.2). A mixed-effects logistic regression on listener responses with trial number as a fixed effect, and including by-pair random slopes and intercepts, showed a significant improvement in accuracy overall, $z = 14.4, p < 0.001$. Accuracy also differed significantly *across* conditions (Fig. 5.2): adding an additional main effect of condition to our logistic model provided a significantly better fit, $\chi^2(2) = 10.8, p = 0.004$. Qualitatively, the *coarse* condition was easiest for participants, the *fine* condition was hardest, and the *mixed* condition was roughly in between. Finally, the (log) response time taken by the speaker to choose an utterance also decreased significantly over the course of the game, $t = -19.7, p < 0.001$, indicating that lexical mappings became increasingly established or accessible.

Partners converge on similar lexica

Another indicator of successful learning is convergence or alignment of lexica across partners in a dyad. Before using post-test responses to compute similarity *across* partners, however, we examine the internal consistency *within* an individual's post-test responses. For each participant, we counted the number of mismatches between the two directions of the lexicon question (e.g. if they clicked the word 'mawa' when we showed them one of the blue squares, but failed to click that same blue square when we showed 'mawa'). In general, participants were quite consistent: out of 128 cells in the lexicon matrix (16 words $\times$ 8 objects), the median number of mismatches was 2 (98% agreement), though the distribution has a long tail (mean $=$ 7.3). We therefore conservatively take a participant's final lexicon to be the *intersection* of their word-to-object and object-to-word responses.

Using these estimates of each participant's lexicon, we compute the overlap between partners. For most pairs, partners aligned strongly by the end, with a median post-test overlap of 97.6% (125 out of 128 entries). Because these matrices were extremely sparse, however, just a a few mismatches could have a large impact on performance. Overall accuracy in the game is strongly correlated with alignment: partners who reported more similar lexica at the end tended to perform better at the task ($r = 0.77$).

Despite these markers of success at the group level, individual performance was bimodal: a subpopulation of 29 games (11% of coarse games, 18% of mixed, and 39% of fine) still showed relatively poor performance, sometimes at chance, by the end of the game. For the subsequent analyses focusing on the content of the lexicon, we exclude games with fourth-quartile accuracy below the pre-registered criterion of 75% to ensure we are examining only successful lexica.

Contextual pressures shape the lexicon

We predicted that in contexts regularly requiring speakers to make fine distinctions among objects at subordinate levels of the hierarchy, we would find lexicalization of specific terms for each object (indeed, a one-to-one mapping may be the most obvious solution in a task with only 8 objects). Conversely, when no such distinctions were required, we expected participants to adaptively lexicalize more abstract terms. One coarse signature of this prediction lies in the *efficiency* of the resulting lexicon: lexicalizing abstract terms should require participants to use fewer terms overall.

To test this prediction, we counted the number of words in each participant's reported lexicon (i.e. the words for which they marked at least one object in the post-test). We found that par-

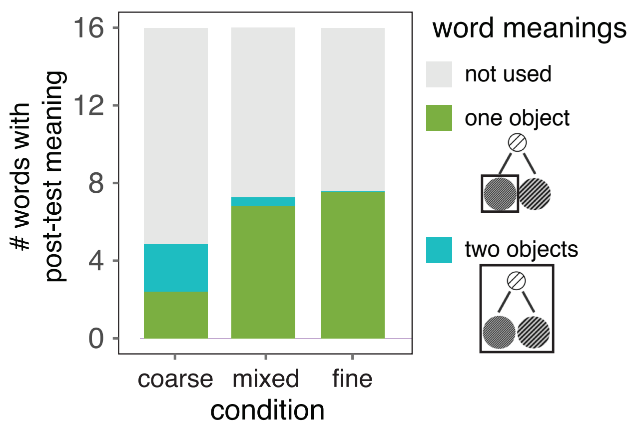

Figure 5.3: Pragmatic demands of context shape the formation of abstractions. Mean number of words participants reported with specific meanings (applying to 1 object) or abstract meanings (applying to 2 objects).

ticipants in the *coarse* condition reported significantly smaller, more efficient lexica ($m = 4.9$ words) than participants in the *mixed* and *fine* conditions ($m = 7.4, t = 10.3, p < 0.001$ and $m = 7.6, t = 9.5, p < 0.001$, respectively; see Fig. 5.3A). At the same time, the smaller lexicon provided equivalent coverage of objects: the median number of objects where participants agreed on the same word or words was 7, 6.5, and 7, respectively.

If participants in the *coarse* condition can get away with fewer words in their lexicon, what are the meanings of the words they do have? We counted the numbers of 'specific' terms (e.g. words that refer to only one object) and 'abstract' terms (e.g. words that refer to two objects) in the post-test. We found that the likelihood of lexicalizing abstractions differed systematically across conditions (see Fig. 5.3). Participants in the *fine* condition reported lexica containing exclusively specific terms, while participants in the *coarse* condition reported significantly more abstract terms ($m = 2.5, p < 0.001$).

These data also reveal an interesting asymmetry in lexicon content across conditions: while abstractions are entirely absent from the *fine* condition, participants in the other conditions often

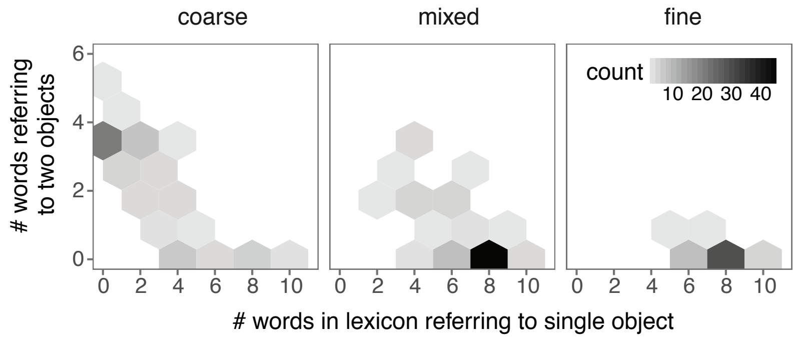

Figure 5.4: Diversity of terms within reported lexica: many participants in the *coarse* condition reported a mixture of abstract and specific terms.

reported a mixture of terms (see Fig. 5.4). In the *coarse* condition, for instance, participants could in principle perform optimally with only four abstract terms and no specific terms. While this was the modal system that emerged (reported in the post-test by nearly 1/3 of participants), the average proportion of abstract (vs. specific) terms *within* each participant's lexicon in the *coarse* condition ($m = 0.56$) was significantly higher than in the other conditions ($p < 0.001$, exploratory).

5.2 Model-based Analysis

Our post-test provides some insight into the end-result of lexicalization under different communicative contexts, but understanding the *dynamics* of lexicalization requires a more detailed analysis of behavioral trajectories. How do lexica shift and develop over the course of interaction?

In this section, we present a statistical model of this progression. We assume that on any given trial, speakers and listeners are rationally producing and interpreting utterances given some internal lexicon, and we use a Bayesian statistical model to infer their lexicon from their behavior. First, this analysis validates our post-test measures of lexical meaning against actual behavioral usage — if participant reports are internally consistent, the model's posterior near the end of the game should predict their post-test responses. Second, we can examine the time-course of lexical emergence by inspecting lexica inferred from early behavior in the game.

5.2.1 Generative model

We begin with a generative model of how agents use their underlying lexicon to produce and interpret language. This model provides a linking function assigning a likelihood to the speaker utterances and listener choices we observe on each trial, given any latent lexicon. We adopt the probabilistic Rational Speech Act (RSA) framework, which has been successful in recent years at capturing a broad array of pragmatic phenomena in language use (N. D. Goodman & Frank, 2016; Franke & Jäger, 2016). This framework captures the Gricean assumption of cooperativity: a pragmatic speaker S_1 attempts to be informative in context while a pragmatic listener L_1 inverts their model of the speaker to infer the intended target. The chain of recursive social reasoning grounds out in a *literal listener* L_0, which directly soft-maximizes its lexicon, $\mathcal{L}^t(w, o)$, to interpret a given utterance. This model can be formally specified as follows:

$$\begin{aligned}
L_0(o_i|w, \mathcal{L}^t) &\propto \exp\{\mathcal{L}^t(w, o_i)\} \\
S_1(w|o_i, \mathcal{L}^t) &\propto \exp\{\ln L_0(o_i|w, \mathcal{L}^t)\} \\
L_1(o_i|w, \mathcal{L}^t) &\propto S_1(w|o_i, \mathcal{L}^t)P(o_i)
\end{aligned}$$

where o_i is a chosen object and w an uttered word.

We use these pragmatic speaker and listener likelihood functions to link latent lexica, represented as a matrix of real values $\ell^t_{w,o} \in \mathbb{R}$, to behavior. This allows us to then use Bayesian inference to back out each participant's effective lexicon from their trial-by-trial behavior. Because each trial has only a single choice for each player, we pool statistics within k epochs of the data (we choose $k = 6$ such that each target appears exactly twice in each epoch). For each epoch, we sample lexical entries

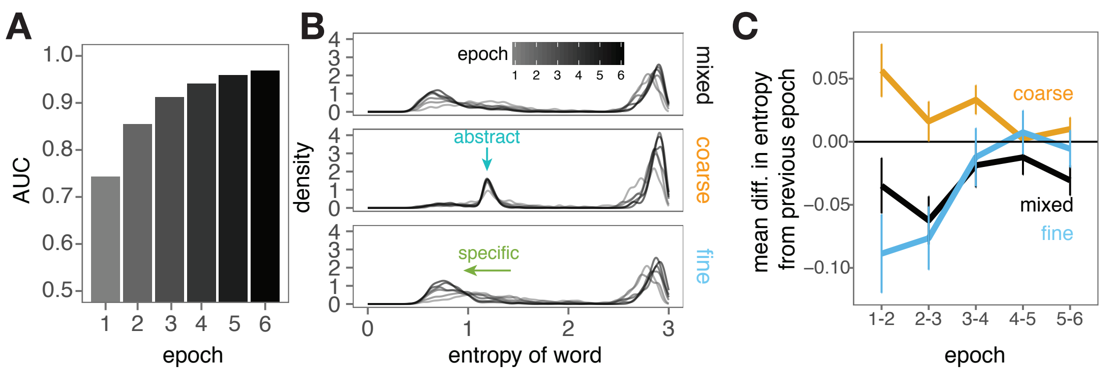

Figure 5.5: Model-based results. (A) A logistic classifier based on inferred lexical entries accurately predicts post-test responses. (B) Entropy of posterior word extensions show coalescence across epochs for each condition. (C) Mean change in entropy at the word level from trial to trial (error bars are ± 1 SE)

from independent Gaussian priors:

$$\ell^k_{o,w} \sim \mathcal{N}(0, 5)$$

This prior is intended to regularize lexicon entries to be relatively close to 0, inducing a bias toward sparsity.

We approximate the posterior of this model separately for each pair using mean-field variational inference, implemented in the probabilistic programming language WebPPL (N. D. Goodman & Stuhlmüller, electronic; Ritchie, Horsfall, & Goodman, 2016). The approximating family for each random variable is Gaussian. We approximate the joint posterior over all lexical entries used in each epoch by each participant.

5.2.2 Validating post-test responses

We begin by showing that the lexical entries we infer for each participant accurately predict their post-test responses. We constructed a logistic classifier from our posterior on each epoch: for each object-word pair (o, w) in the post-test response matrix, we computed the marginal posterior probability $P(\ell_{o,w} > 0.5|\theta_{o,w})$, where $\theta_{o,w}$ are the corresponding variational parameters (i.e. the mean and variance of the approximating Gaussian). This gives the posterior probability that word w applies to object o. We evaluated the performance of this classifier by constructing an ROC curve that shows the tradeoff between hits and false alarms as the discrimination criterion is varied. We found that the classifier based on the final epoch predicts post-test responses with excellent accuracy (AUC: 0.98; see Fig. 5.5A). This indicates that the post-test lexicon is indeed linked to behavior as predicted by RSA, validating both the post-test measure and the results of our Bayesian analysis.

Furthermore, we found that the corresponding posterior predictives from earlier epochs predicted final post-test responses less well, even though they were learned from the same number and type of behavioral observations (Fig. 5.5A). Still, even the classifier based on the earliest epoch performs above chance, indicating that some information about the final lexicon is available from the earliest trials. These patterns are suggestive of a path-dependent process where the lexicon gradually coalesces from initially arbitrary associations over the course of interaction. We next turn to the earliest stages of this process.

5.2.3 Examining early time course

One advantage of the statistical approach we develop here is the ability to make descriptive inferences about the meanings being used in settings where we *don't* ask participants for explicit judgements—in particular, in early trials of our games.

Our primary measure of interest is the *entropy* of the extension of words over the eight objects. The entropy of a particular word is near zero when its meaning is peaked on a single object, and is maximized when could apply equally to all objects (e.g. for a novel word that has not yet been used). We expect abstract terms to lie in between these extremes. We obtain the extension distribution for each word by running it through our L_0 model, essentially asking how likely it is to refer to each of the eight objects.[‡] We use the MAP estimate of the lexicon. The resulting distribution of estimated

[‡]Using L_0, rather than L_1 or $\mathcal{L}$, gives us a notion of word extension that is close to the underlying lexi-

word entropies, aggregated for each epoch and condition, is shown in Fig. 5.5B. Abstract terms begin to form early (epoch 2) in the *coarse* condition, and remain stable throughout the game. In contrast, specific terms are relatively slow-forming (epoch 4-5) in the other two conditions. The peak near an entropy of 3 reflects the inferred ambiguity of words that were not used or used randomly.

Because these distributions are aggregated across words, however, they leave open the possibility that lexica are not stabilizing or coalescing but simply cycling through different words each epoch. We address these dynamics more thoroughly at the *word* level by computing the difference in each word's entropy from epoch to epoch (Fig. 5.5C). For all conditions, we found that the entropy of individual words changed less over later epochs (i.e. the difference scores approached zero), indicating that meanings gradually stabilized. There are also differences across conditions: words in the *mixed* and *fine* condition began with high entropy reduction (becoming more specific) which continued through the final epochs, while words in the *coarse* condition actually seemed to increase in entropy across the game on average.

These preliminary results, then, may reflect a combination of narrowing and broadening depending on condition. Unknown words can initially refer to any of the objects and only acquire more informative meanings as agents learn through interaction. Yet in the coarse condition where agents are quick to adopt meanings, the rest of the game may be spent paring down the lexicon instead.

5.3 Discussion

How and why do abstractions emerge in local interactions? We hypothesized that although communicative contexts requiring fine distinctions would favor one-to-one object-word mappings, pressures for efficiency would allow abstractions to emerge in coarser contexts. By manipulating context statistics in a real-time experiment, we found evidence for these pragmatic influences on interactive

con while influenced by non-identifiability of parameters. For instance, $\mathcal{L}$ has an overall scaling per row that doesn't influence behavior.

convention formation.

Our results may help to illuminate the relationship between our concepts and words, which are often treated interchangeably. While our mental taxonomies are adaptive to the natural perceptual structure of the world (Mervis & Rosch, 1981) it is far from inevitable that all levels of these conceptual hierarchies become conventionalized as lexical items. There are many perfectly natural concepts that are not represented by distinct words in the English language: for instance, we do not have words for each tree in our yards, or for ad-hoc concepts (Barsalou, 1983). Indeed, English speakers are often fascinated by foreign words like the Danish "hygge" (a specific notion of coziness) or Scottish "tartle" (hesitating when introducing someone because you've forgotten their name) that are difficult to express in English. Our results highlight communicative needs to distinguish, in context, as a force behind the choice to lexicalize some fine-grained concepts. A related direction for future work is to explore the relationship between communicative need and *basic-level* structure.

While we showed how abstract words emerge from efficiency even in a task requiring only reference to individual objects, there are other clear functional advantages to having abstract terms in the lexicon. For one, they allow speakers to efficiently refer to large, potentially infinite, sets of things, and make generalizations about categories, e.g. "Dogs bark" (Tessler & Goodman, 2019). Future work should explore this as an additional pressure toward abstract, nested nouns. Similarly, the option to refer to more specific concepts with compound terms (e.g. "spotted dog"), which was not available in our experiment, may impact final conventions. We expect that labels will become lexicalized when the cost incurred by frequently using a compositional construction exceeds the cost of adding an additional word to the lexicon. Future work should also explore these hypotheses about how lexicalization of nominal terms trades off with compositionality.

Finally, although we implemented a purely statistical Bayesian data analysis model to infer lexica, it is also possible to consider a cognitive model of participants' own lexical inferences. Indeed, our findings are consistent with a recent cognitive model of convention formation which explained the rapid coordination on efficient but informative lexical terms as a process of mutual lexical learning (Hawkins et al., 2017). In this model, each agent assumes their partner is rationally producing cooperative utterances under some latent lexicon; given initial uncertainty over the contents of that lexicon, agents can invert their model of their partner to infer their lexicon from observable behavior. The different dynamics we observed across conditions, then, may be the consequence of different lexical inferences in different local contexts. Further, while we used RSA as a linking function in our statistical model, a cognitive model would allow us to test to what extent pragmatic reasoning is necessary to explain behavior.

Our shared lexical conventions are richly structured systems with meanings at multiple levels of abstraction. There is now abundant evidence that languages adapt to the needs of their users, and the context-sensitive emergence of abstractions demonstrated in this paper suggests that the driver of this adaptation may lie in the remarkably rapid adaptability of agents themselves. We are constantly supplementing our existing language with local conventions as we need them. Our separate minds may organize the world into meaningful conceptual hierarchies but our shared language only evolves to reflect this structure when it is communicatively relevant.

6

Disentangling contributions of visual information and interaction history in the formation of graphical conventions

From ancient etchings on cave walls to modern digital displays, visual communication lies at the heart of key human innovations (e.g., cartography, data visualization) and forms a durable foundation for the cultural transmission of knowledge and higher-level reasoning. Perhaps the most basic and versatile technique supporting visual communication is drawing, the earliest examples of which date to at least 40,000-60,000 years ago (Hoffmann et al., 2018). What began as simple mark mak-

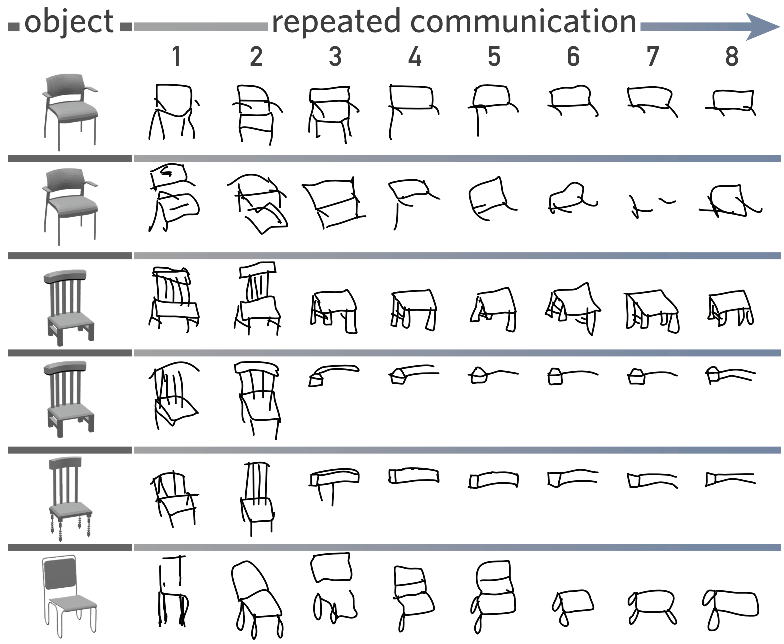

Figure 6.1: Repeated visual communication depicting the same object.

ing has since been adapted to a wide array of applications, ranging from photorealistic rendering to schematic diagrams consisting entirely of symbols.

Even in the relatively straightforward case of drawing from life, there are countless ways to depict the same object. How does a communication medium spanning such a broad range of appearances reliably convey meaning? On the one hand, prior work has found that semantic information in a figurative drawing, i.e., the object it represents, can be derived purely from its visual properties (Fan, Yamins, & Turk-Browne, 2018). On the other hand, other work has emphasized the role of socially-mediated information for making appropriate inferences about what even a figurative drawing represents (N. Goodman, 1976).

How can these two perspectives be reconciled? Our approach is to consider the joint contributions of visual information and social context in determining how drawings derive meaning (Abell, 2009), and to propose that a critical factor affecting the balance between the two may be the amount of shared knowledge between communicators. Specifically, we explore the hypothesis that accumulation of shared knowledge via extended visual communication may promote the development of increasingly schematic yet effective ways of depicting an object, even as these *ad hoc* graphical conventions may be less readily apprehended by others who lack this shared knowledge.

To investigate this hypothesis, we used an interactive drawing-based reference game in which two participants repeatedly communicated about visual objects. We examined both how their task performance and the drawings they produced changed over time (see Fig. 6.1). Our approach was inspired by a large literature that has explored how extended interaction influences communicative behavior in several modalities, including language (H. H. Clark & Wilkes-Gibbs, 1986; Hawkins et al., 2017), gesture (Goldin-Meadow, McNeill, & Singleton, 1996), and drawings (Garrod et al., 2007; Galantucci, 2005). There are three aspects of the current work that advance our prior understanding: *first*, we include a control set of objects that were not repeatedly drawn but only shown at the beginning and end of the interaction, allowing us to measure the specific contribution of repeated reference vs. general practice effects; *second*, we measure how strongly the visual properties of drawings drive recognition in the absence of interaction history for naive viewers, while equating other task variables; and *third*, we employ recent advances in computer vision to quantitatively characterize changes in the high-level visual properties of drawings across repetitions.

6.1 How does repeated reference support successful visual communication?

Our first goal was to understand how people learn to communicate about visual objects across repeated visual communication. To accomplish this, we developed a drawing-based reference game for two participants. On each trial, both participants shared a *communicative context*, represented by an array of four objects. One of these objects was privately designated the 'target' to the sketcher. The sketcher's goal was to draw the target so that the viewer could select it from the array as quickly

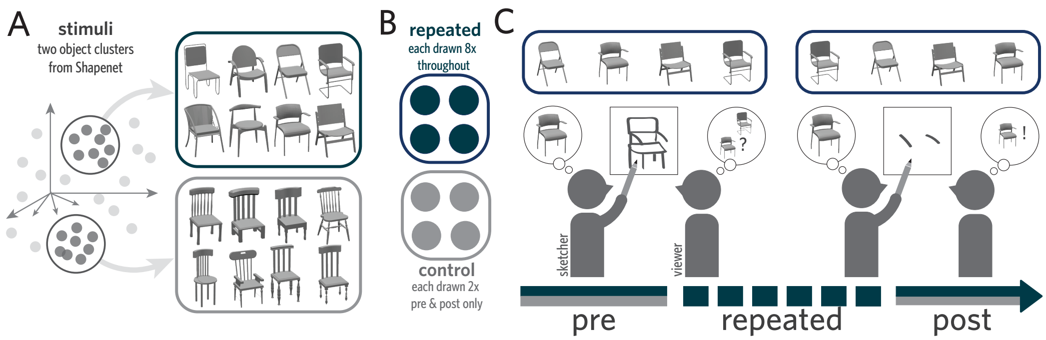

Figure 6.2: (A) Stimuli from ShapeNet. (B) Each pair of participants was randomly assigned two sets of four objects, each set from one of the two categories. (C) Repeated objects drawn eight times throughout; control objects drawn once at the beginning and end of each interaction.

and accurately as possible. We hypothesized that learning would be *object-specific*: that over repeated visual reference to a particular object, participants would discover ways of depicting that object more effectively relative to non-repeated control objects.

6.1.1 Methods: Visual communication experiment

Participants

We recruited 138 participants from Amazon Mechanical Turk, who were grouped into 69 pairs (Hawkins, 2015). Within each experimental session, one participant was assigned the sketcher role and the other the viewer role, and these role assignments remained the same throughout the experiment. Data from two pairs were excluded due to unusually low performance (i.e., accuracy < 3 s.d. below the mean). In this and subsequent experiments, participants provided informed consent in accordance with the Stanford IRB.

Stimuli

In order to make our task sufficiently challenging, we sought to construct communicative contexts consisting of objects whose members were both geometrically complex and visually similar. To accomplish this, we sampled objects from the ShapeNet (Chang et al., 2015), a database containing a large number of 3D mesh models of real-world objects. We restricted our search to 3096 objects belonging to the `chair` class, which is among the most diverse and abundant in ShapeNet. To identify groups of visually similar chairs, we first extracted high-level visual features from 2D renderings of each object using a deep convolutional neural network (DCNN) architecture, VGG-19 (Simonyan & Zisserman, 2014). This network had been previously trained to recognize objects in photos from the ImageNet database (Deng et al., 2009), containing 1.2 million natural photographs of 1000 different object classes. Trained DCNN models have been shown to predict human perceptual similarity judgments about objects (Kubilius, Bracci, & de Beeck, 2016; Peterson, Abbott, & Griffiths, 2018), as well as neural population responses in visual cortex during object recognition (Yamins et al., 2014; Güçlü & van Gerven, 2015). As such, they provide a principled choice of encoding model for extracting high-level visual information from images. Following previous work that has employed DCNN models to evaluate perceptual similarity (Peterson et al., 2018; Kubilius et al., 2016), for each image we extract a 4096-dimensional feature vector reflecting activations in the second fully-connected layer (i.e., `fc6`) of VGG-19, a higher layer in the network. We then applied dimensionality reduction (PCA) and k-means clustering on these feature vectors, yielding 70 clusters containing between 2 and 80 objects each. Among clusters that contained at least eight objects, we manually identified two visual categories containing eight objects each (Fig. 6.2A).

Task Procedure

On each trial, both participants were shown the same set of four objects in randomized locations. One of the four objects was highlighted on the sketcher's screen to designate it as the target. Sketchers drew using their mouse cursor in black ink on a digital canvas embedded in their web browser (300×300 pixels; pen width = 5px). Each stroke was rendered on the viewer's screen in real time and sketchers could not delete previous strokes. The viewer aimed to click one of the four objects as soon as they were confident of the identity of the target, and participants received immediate feedback: the sketcher learned when and which object the viewer had clicked, and the viewer learned the true identity of the target. Both participants were incentivized to perform both quickly and accurately. They both earned an accuracy bonus for each correct response, and the sketcher was required to complete their drawings in 30 seconds or less. If the viewer responded correctly within this time limit, participants also received a speed bonus inversely proportional to the time taken until the response.

Design

For each pair of participants, two sets of four objects were randomly sampled to serve as communication contexts: one was designated the *repeated* set while the other served as the *control* set (Fig. 6.2B).* The experiment consisted of three phases (Fig. 6.2C). During the repeated reference phase, there were six repetition blocks of four trials, and each of the four *repeated* objects appeared as the target once in each repetition block. In a pretest phase at the beginning of the experiment and a posttest phase at the end, both repeated and control objects appeared once as targets (in their respective contexts) in a randomly interleaved order.

6.1.2 Results

Because objects were randomly assigned to repeated and control conditions, we expected no differences in task performance in the pretest phase. We found that pairs identified the target at rates well above chance in this phase (75.7% repeated, 76.1% control, chance = 25%), suggesting that they were engaged with the task but not at ceiling performance. We found no difference in accuracy across conditions (mean difference: 0.3%, bootstrapped CI: $[-7\%, 7\%]$).

*In half of the pairs, the four control objects were from the same stimulus cluster as repeated objects; in the other half, they were from different clusters. The rationale for this was to support investigation of between-cluster generalization in future analyses. In current analyses, we collapse across these groups.

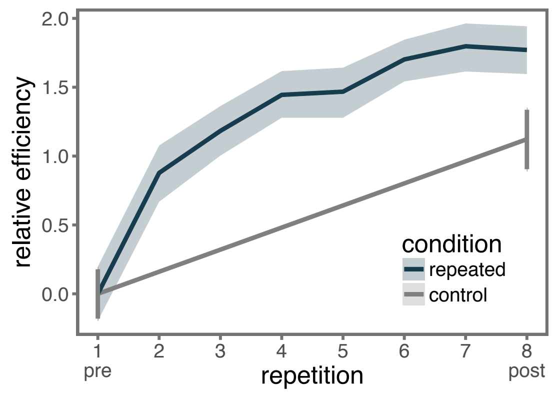

Figure 6.3: Communication efficiency across repetitions. Efficiency combines both speed and accuracy, and is plotted relative to the first repetition. Error ribbons represent 95% CI.

In order to measure how well pairs learned to communicate throughout the rest of their interaction, we used a measure of communicative efficiency (the *balanced integration score,* Liesefeld & Janczyk, 2019) that takes both accuracy (i.e., proportion of correct viewer responses) and response time (i.e., latency before viewer response) into account. This efficiency score is computed by first z-scoring accuracy and response time across repetitions within an interaction to map values from different interactions to the same scale, and then subtracting the standardized response time from standardized accuracy. It is highest when pairs are both fast and accurate, and lowest when they make more errors and take longer, relative to their own performance on other trials.

To evaluate changes in communicative efficiency, we fit a linear mixed-effects model including random intercepts, slopes, and interactions for each pair of participants. We found a main effect of increasing communicative efficiency for all targets between the *pre* and *post* phases ($b = 1.45,\ t = 14.3,\ p < 0.001$), reflecting general improvements due to task practice. Critically, however, this analysis also revealed a reliable interaction between phase and condition: communicative efficiency improved to a greater extent for repeated objects than control objects ($b = 0.648,\ t = 3.09,\ p = 0.003$; see Fig. 6.3). Analysis of changes in raw accuracy yielded a similar result: performance on repeated objects improved by 14.5%, while performance on control objects only improved by 7.1%. Together, these data show that there are benefits of repeatedly communicating about an object that accrue specifically to that object, suggesting the formation of object-specific graphical conventions.

6.2 What explains gains in efficiency?

Our visual communication experiment established that pairs of participants coordinate on more efficient and *object-specific* ways of depicting targets. This raises the question: to what extent do these gains in efficiency reflect the accumulation of *interaction-specific* shared knowledge between a sketcher and viewer, as opposed to the combination of task practice and the inherent visual properties of their drawings?

To disentangle the contributions of these different factors, we conducted two control experiments to estimate the how recognizable these drawings were to naive viewers outside the social context in which they were produced. Participants in one control group were shown a sequence of drawings taken from a single interaction, closely matching the experience of viewers in the communication experiment. Participants in a second control group were instead shown a sequence of drawings pieced together from many different interactions, thus disrupting the continuity experienced by viewers paired with a single sketcher. Insofar as interaction-specific shared knowledge contributed to the efficiency gains observed previously, we hypothesized that the second group would not improve as much over the course of the experimental session as the first group would.

6.2.1 Methods: Recognition Control Experiments

Participants

We recruited 245 participants via Amazon Mechanical Turk. We excluded data from 22 participants who did not meet our inclusion criterion for accurate and consistent response on attention-check

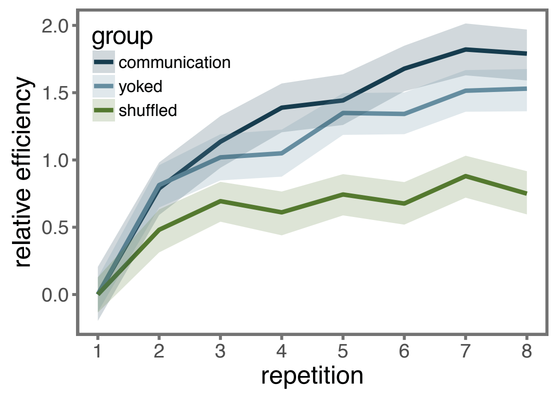

Figure 6.4: Comparing drawing recognition performance between viewers in communication experiment with those of yoked and shuffled control groups. Error ribbons represent 95% CI.

trials (see below).

Task, Design, & Procedure

On each trial, participants were presented with a drawing and the same set of four objects that accompanied that drawing in the original visual communication experiment. They also received the same accuracy and speed bonuses as viewers in the communication experiment. To ensure task engagement, we included five identical attention-check trials that appeared once every eight trials. Each attention-check trial presented the same set of objects and drawing, which we identified during piloting as the most consistently and accurately recognized by naive participants. Only participants who responded correctly on at least four out of five of these trials were retained in subsequent analyses.

Each participant was randomly assigned to one of two conditions: a *yoked* group and a *shuffled* group. Each yoked participant was matched with a single interaction from the original cohort and viewed 40 drawings in the same sequence the original viewer had. Those in the shuffled group were matched with a random sample of 10 distinct interactions from the original cohort and viewed four drawings from each in turn, which appeared within the same repetition block as they had originally.

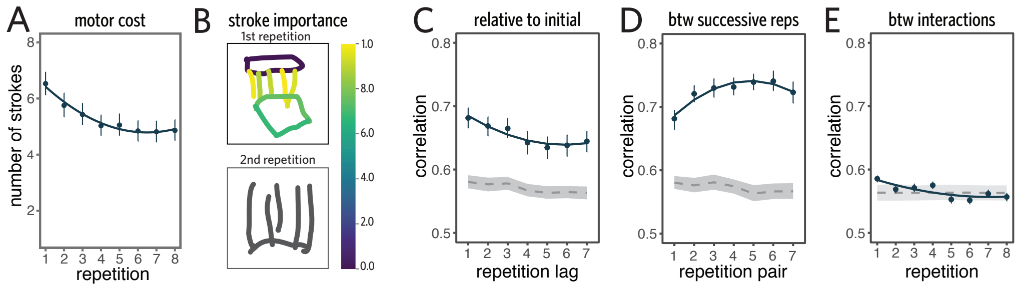

Figure 6.5: (A) Sketchers use fewer strokes over time. (B) Visualizing importance of individual strokes in successive drawings. (C) Drawings become increasingly dissimilar from initial drawing. (D) Drawings become more consistent from repetition to repetition. (E) The same object is drawn increasingly dissimilarly by different sketchers. Error ribbons represent 95% CI, dotted lines represent permuted baseline.

For example, if a drawing was produced in the fifth repetition block in the original experiment, then it also appeared in the fifth block for shuffled participants.

At the trial level, groups in both conditions thus received exactly the same visual information and performed the task under the same incentives to respond quickly and accurately. At the repetition level, both groups received exactly the same amount of practice recognizing drawings. Thus any differences between these groups are attributable to whether drawings came from the same communicative interaction, which would support the accumulation of interaction-specific experience, or from several different interactions, where such accumulation would be minimal.

6.2.2 Results

Interaction-specific history enhances recognition by third-party observers

We compared the yoked and shuffled groups by measuring changes in recognition performance across successive repetitions using the same efficiency metric we previously used. We estimated the magnitude of these changes by fitting a linear mixed-effects model that included group (yoked vs. shuffled), repetition number (i.e., first through eighth), and their interaction, as well as random intercepts and slopes for each participant. While we found a significant increase in recognition performance across both groups ($b = 0.18,\ t = 12.8,\ p < 0.001$), we also found a large and reliable interaction: yoked participants improved to a substantially greater degree than shuffled participants ($b = 0.10,\ t = 4.9,\ p < 0.001$; Fig. 4). Examining accuracy alone yielded similar results: the yoked group improved to a greater degree across the session (yoked: +15.8%, shuffled: +5.6%). Taken together, these results suggest that third-party observers in the yoked condition who viewed drawings from a single interaction were able to take advantage of this continuity to more accurately identify what successive drawings represented. While observers in the shuffled condition still improved over time, being deprived of this interaction continuity made it relatively more difficult to interpret later drawings.

Viewer feedback also contributes to gains in performance

Unlike viewers in the interactive visual communication experiment, participants in the yoked condition made their decision based only on the whole drawing and were unable to interrupt or await additional information if they were still uncertain. Sketchers could have used this feedback to modify their drawings on subsequent repetitions. As such, comparing the yoked and original communication groups provides an estimate of the contribution of these viewer feedback channels to gains in performance (Schober & Clark, 1989). In a mixed-effects model with random intercepts, slopes, and interactions for each unique trial sequence, we found a strong main effect of repetition ($b = 0.23,\ t = 12.8,\ p < 0.001$), as well as a weaker but reliable interaction with group membership ($b = -0.05,\ t = -2.2,\ p = 0.032$, Fig. 6.4), showing that the yoked group improved at a more modest rate than viewers in the original communication experiment had.

To better understand this interaction, we further examined changes in the accuracy and response time components of the efficiency score. We found that while viewers in the communication experiment were more accurate than yoked participants overall (communication: 88%, yoked: 75%), *improvements* in accuracy over the course of the experiment were similar in both groups (communication: +14.5%, yoked: +15.8%). The interaction instead appeared to be driven by differential reductions in response time between the first and final repetitions (communication: 10.9s to 5.84s; yoked: 4.66s to 3.31s). These reductions were smaller in the yoked group, given that these participants did not need to wait for each stroke to appear before making a decision, and thus may have already been closer to floor.

6.3 How do visual features of drawings change over the course of an interaction?

The results so far show that repeated visual communication establishes object-specific, interaction-specific ways of efficiently referring to objects. An intriguing implication is that interacting pairs achieved this by gradually forming *ad hoc* graphical conventions about what was relevant and sufficient to include in a drawing to support rapid identification of the target object. Here we explore this possibility by examining how the drawings themselves changed throughout an interaction. Concretely, we investigated four aspects that would reflect the increasing contribution of interaction-specific shared knowledge: *first*, decreasing number of strokes used (i.e., reducing motor cost of each drawing); *second*, increasing dissimilarity from the initial drawing produced (i.e., cumulative drift from the starting point); *third*, increasing similarity between successive drawings (i.e., convergence on internally consistent ways of depicting objects within an interaction); *fourth*, increasing dissimilarity between drawings of the same object produced in different interactions (i.e., discovery of multiple viable solutions to the coordination problem).

6.3.1 Measuring visual similarity between drawings

Measuring visual similarity between drawings depends upon a principled approach for encoding their high-level visual properties. Here we capitalize on recent work validating the use of deep convolutional neural network models to encode such perceptual content in drawings (Fan et al., 2018). As when identifying clusters of similar object stimuli, we again used VGG-19 to extract 4096-dimensional feature vector representations for drawings of every object, in every repetition, from every interaction. Using this feature basis, we compute the similarity between any two drawings as the Pearson correlation between their feature vectors (i.e., $s_{ij} = {cov(\vec{r}_i, \vec{r}_j)}/{\sqrt{var(\vec{r}_i) \cdot var(\vec{r}_j)}}$).

6.3.2 Results

Fewer strokes across repetitions

A straightforward explanation for the gains in communication efficiency observed in Part I is that sketchers were able to use fewer strokes per drawing to achieve the same level of viewer recognition accuracy. Indeed, we found that the number of strokes in drawings of repeated objects decreased steadily as a function of repetition in a mixed-effects model ($b = -0.216,\ t = -6.00$; Fig. 6.5A), suggesting that pairs were increasingly able to rely upon shared knowledge to communicate efficiently. This result raises a question about *which* strokes are preserved across successive repetitions during the formation of graphical conventions. In ongoing work, we are using a lesion method to investigate the "importance" of each stroke within a drawing for explaining similarity to the next repetition's drawing of that object. We re-render the drawing without each stroke and compute the similarity, yielding a heat map across strokes (see Fig. 6.5B for an example visualization). The more dissimilar the lesioned drawing without a particular stroke is to an intact version of the next repetition's drawing, the more "important" we consider that stroke to be.

Increasing dissimilarity from initial drawing

Mirroring the observed reduction in the number of strokes across repetitions, we hypothesized that there was also cumulative change in the visual content of drawings across repetitions. Concretely, we predicted that drawings would become increasingly dissimilar from the initial depiction. We tested this prediction in a mixed-effects regression model including linear and quadratic terms for repetition as well as intercepts for each target and pair. We found a significant decrease in similarity to the initial round across successive repetitions, ($b = -0.62,\ t = -5.59$; Fig. 6.5C), suggesting that later drawings had moved to a different region of visual feature space. However, since the entire distribution of drawings may have drifted to a different region of the visual feature space for generic reasons (i.e., because they were sparser overall), we conducted a stricter permutation test. We scrambled drawings across pairs but within each repetition and target and re-ran our mixed-effects model. The observed effect fell outside this null distribution ($CI = [-3.53 - 0.88],\ p < .001$), showing that successive drawings by the same sketcher deviated from their own initial drawing to a greater degree than would be expected due to generic differences between drawings made at different timepoints in an interaction.

Increasing internal consistency within interaction

As sketchers modified their drawings across successive repetitions, we additionally hypothesized that they would converge on increasingly consistent ways of depicting each object. To test this prediction, we computed the similarity of successive drawings of the same object made in the same interaction (i.e. repetition k to $k + 1$). A mixed-effects model with random intercepts for both object and pair showed that similarity between successive drawings increased substantially throughout an interaction ($b = 0.53,\ t = 5.03$; Fig. 6.5). Again, we compared our empirical estimate of the magnitude of this trend to a null distribution of slope t values generated by scrambling drawings across pairs. The observed increase fell outside this null distribution ($CI = [-3.21, -0.60], p < .001$), providing evidence that increasingly consistent ways of drawing each object manifested only for series of drawings produced within the same interaction.

Increasingly different drawings across interactions

Our recognition control experiments suggested that the graphical conventions discovered by different pairs were increasingly opaque to outside observers. This effect could arise if early drawings were more strongly constrained by the visual properties of a shared target object, but later drawings diverged as different pairs discovered different equilibria in the space of viable graphical conventions. Under this account, drawings of the same object from different pairs would become increasingly dissimilar from each other across repetitions. We tested this prediction by computing the mean pairwise similarity between drawings of the same object within each repetition index, but produced in different interactions. Specifically, for each object, we considered all interactions in which that object was repeatedly drawn. Then, for each repetition index, we computed the average similarity between drawings of that object. In a mixed-effects regression model including linear and quadratic terms, as well as random slopes and intercepts for object and pair, we found a small but reliable negative effect of repetition on between-interaction drawing similarity ($b = -1.4,\ t = -2.5$; Fig. 6.5E). We again conducted a permutation test to compare this t value with what would be expected from scrambling sketches across repetitions for each sketcher and target object. We found that the observed slope was highly unlikely under this distribution ($CI = [-0.57, 0.60],\ p < 0.001$), even if the similarity at each round was not so unlikely.

6.4 Discussion

In this paper, we investigated the joint contributions of visual information and social context to determining the meaning of drawings. We observed in an interactive Pictionary-style communication game that pairs of participants discover increasingly sparse yet effective ways of depicting objects over repeated reference. Through a series of control experiments, we demonstrated that these conventionalized representations were both object-specific and interaction-specific: drawings were harder for independent viewers to recognize without sharing the same interaction history. Furthermore, by analyzing the high-level visual features of drawings, we found that they became increasingly consistent within an interaction, but that different pairs discover different equilibria in the space of viable graphical conventions. Taken together, our findings suggest that repeated visual communication promotes the emergence of depictions whose meanings are increasingly determined by interaction history rather than their visual properties alone.

A key experimental design choice was the use of visual objects as the targets of reference, by contrast with the verbal labels or audio clips used in prior work (Galantucci & Garrod, 2011; Fay et al., 2010). As such, communication between the sketcher and viewer was grounded in the same visual information about the appearance of these objects, encouraging the production of more 'iconic' initial drawings that more strongly resembled the target object (Verhoef, Kirby, & de Boer, 2016; Perlman, Dale, & Lupyan, 2015). As their communication became increasingly efficient across repetitions, their drawings became simpler and apparently more 'abstract'. An exciting direction for future work is to develop robust and principled computational measures of the degree of visual correspondence between any drawing and any target object, thereby shedding light on the nature of visual abstraction and iconicity.

A second important design choice was the use of a speed bonus incentivizing participants to complete trials quickly. What role do such incentives play in the formation of graphical conventions? Recent computational models of visual communication have found that both how costly a drawing is to produce (i.e., time/ink) and how informative a drawing is in context are critical for explaining the way people spontaneously adjust the level of detail to include in their drawings in one-shot visual communication tasks (Fan, Hawkins, Wu, & Goodman, 2019). The consequences of this intrinsic preference for less costly drawings may be compounded across repetitions, as the accumulation of interaction history allows people to be equally informative with fewer strokes (Hawkins et al., 2017). The magnitude of these intrinsic costs may vary across individuals, however, and the speed bonus made them explicit.

A major open question raised by our work concerns how people decide what information to preserve or discard across repetitions. One possibility is that successful viewer comprehension is attributed to the most recent strokes produced, leading these to be more strongly preserved. For example, if the viewer was able to correctly identify the target only after the backrest was drawn, the sketcher may continue to selectively draw this part. Another possibility is that sketchers preserve what they judge to be the most diagnostic information about the target, regardless of when the viewer made their response. For example, sketchers may focus on drawing the backrest if it strongly distinguishes the target from distractors in context. Future work should disentangle these possibilities empirically and via development of computational models of visual communication that can learn from task-related feedback, as well as judge which strokes would be most diagnostic.

7

Conclusion

In this dissertation, we considered the computational challenge faced by agents trying to communicate in a variable and non-stationary landscape of meaning. We first proposed a hierarchical Bayesian approach to semantic adaption that formalized three broad mechanisms exposed by the prior literature, and showed that a neural network instantiation of this model successfully adapted to human speakers in an interactive natural-language reference game task. Because theories of adaption have been under-constrained by fine-grained empirical data, we then collected and used recent NLP techniques to analyze a large corpus from a replication of the classic tangrams task. Inspired by insights from these analyses, we designed a targeted artificial language experiment to test our hypothesis that context shapes conventions and a graphical communication experiment to test the object- and interaction-specificity of conventions. We conclude by discussing some broader questions raised by the theoretical perspective we have advanced here.

Generalization in language acquisition Throughout our work, we have assumed a discourse-level structure to an agent's priors. We assume there is uncertainty over how words are used *in the given conversation, by the current partner*. However, there is a broader debate over the timescales at which lexicons and lexicon learning mechanisms operate. In particular, these hierarchical learning mechanisms suggest the possibility of a developmental parallel. Are the lexical learning mechanisms adults use to coordinate on local conventions *within* an interaction the same as those supporting language-learning more broadly?

Most laboratory tasks investigating cross-situational word learning only use a single speaker, and even sophisticated models of cross-situational word learning that account for pragmatic reasoning about speaker intentions (Frank, Goodman, & Tenenbaum, 2009, e.g.) tend to collapse over *who* is talking. Yet, as we have argued throughout this work, there is substantial variability across different speakers. If the majority of child-directed speech only comes from a single primary caregiver, then the child may face a difficult generalization problem once they begin interacting with others. Upon hearing an unfamiliar word from a novel speaker, or a familiar word utterance with an unfamiliar meaning, it could be a quirk of that particular speaker *or* indicative of a globally shared convention. There may therefore be substantial path-dependence in acquisition, as children develop their lexical prior and become better attuned to the overall variability in the population (see E. V. Clark, 2009, Chap. 6).

This slow-developing lexical prior is one of several explanation for why young children are so terrible at coordinating on local conventions in repeated reference games (Glucksberg, Krauss, & Weisberg, 1966; Krauss & Glucksberg, 1977). When an experimenter feeds them the messages that adult speakers produced naturally, they had no trouble, even as they reduced down to one- or two-word utterances. When they played with one another, however, Kindergardeners continued to make errors even after 15-16 repetitions; children as old as fifth grade only improved with assistance from the experimenter and never approached the perfect levels of adult performance. Instead of beginning with the long indefinite descriptions full of hedges and modifiers that adults provide, nursery-school speakers began with short, highly idiosyncratic descriptions like *Mother's dress*. If adult speakers' long hedge-filled messages are indeed motivated by lexical uncertainty, then perhaps young children have simply not obtained enough linguistic variability to calibrate their lexical prior. Alternatively, if the pragmatic reasoning required to produce informative utterance depends on theory of mind, then the high processing demands of the task may simply be inhibited performance (e.g. Setoh, Scott, & Baillargeon, 2016). This remains an under-explored puzzle for future developmental research.

Similarities and differences across communication modalities

> The oral modality is not well suited to conveying messages mimetically (i.e., iconically), even though that function is also important to human languages. This function is, however, very well served by the manual modality. (Goldin-Meadow & McNeill, 1999, p.155)

More broadly, putting these results together with results from previous chapters raises a critical question about the role played by communication modality in convention formation. While linguistic communication is powerful and prevalent, research on the dynamics of adaptation in other communication modalities, including graphical and gestural modalities, is important in several ways. First, it is a core claim of our hierarchical learning model that the mechanisms underlying adaptation and convention formation are domain-general. In other words, there's nothing special about spoken or written language; any ad hoc system that we use to communicate and coordinate with other minds should display similar learning dynamics because they are all trying to coordinate on meaning.

Second, because this hierarchical learning model claims a critical role for the global priors we build up across many interactions with many individuals, we predict that different communication modalities should nevertheless display certain systematic differences in their dynamics. For example, consider our reference game from Chapter 4 where the targets are complex, abstract geometric shapes like tangrams. In the verbal modality, these shapes are highly innominate – we don't have much experience naming or describing them with words, thus our global prior is rather weak and we expect local adaptation to play a much bigger role. In the graphical modality, where you must communicate by drawing on a sketchpad, on the other hand, agents have a much stronger prior rooted in assumptions about shared perceptual systems and visual similarity (though see Fan et al., 2018): explaining these similarity judgements poses its own challenges). Other stimuli have precisely the opposite property: to distinguish between natural images of dogs, for instance, we may have very strong priors in the linguistic modality (e.g. 'husky', 'poodle', 'pug', etc) but drawing the necessary fine distinctions in the graphical modality may be initially very costly, encouraging the formation of local conventions.

Practically speaking, then, considering repeated reference games across different modalities is necessary to (1) test which adaptation effects, if any, are robust & attributable to general mechanisms and (2) explain variance across settings where global priors and local adaptation trade off in different ways. If we adhered solely to the verbal modality, we would be limited to a fairly narrow range of stimuli (e.g. abstract shapes/tangrams) where behavior in the lab isn't totally dominated by strong prior conventions people bring into the interaction. The clearest analogs to repeated linguistic reference games in the style of (Krauss & Weinheimer, 1964) are Pictionary games like this ones used in this chapter. where participants were given a whiteboard to draw on instead of an auditory channel to talk through (Theisen, Oberlander, & Kirby, 2010). For example, Garrod et al. (2007) used a set of 12 concept words with intuitively uncertain graphical priors as targets ("Robert de Niro", "poverty").

Another modality-based manipulation is to attempt to destroy or scramble any meaningful priors that people might carry into the social interaction. For example, Galantucci (2005) introduced a novel 'seismograph' interface for communication – a stylus that could be moved side-to-side or lifted up or down to make contact with the sketch pad while the vertical dimension drifted downward at a constant rate. The resulting messages consequently look nothing like the usual kinds of symbols people create: the relationship between motor actions and perceptual output is broken such that executing a familiar movement for a symbol or numeral instead produces an odd, wavy scribble. Despite the relative lack of priors on signal meanings in this medium, people were nevertheless able to converge on successful signaling systems in repeated reference games (Roberts & Galantucci, 2012; Roberts, Lewandowski, & Galantucci, 2015). Other novel modalities used in iterated reference games include a 'whistle' language where movements along a vertical touch bar slider correspond to changes in pitch (Verhoef, Roberts, & Dingemanse, 2015) and a visual analog where movements along the slider were presented visually (Verhoef, Walker, & Marghetis, 2016). A unified model of coordination in communication ought to be able to account for production and comprehension across these modalities simply by exchanging its encoder and decoder components.

Mechanisms for adaptation While we have focused on how participants coordinate on *lexical* meaning, this is only one of many levels at which conventions may form. In more complex circumstances, there is often initial uncertainty not just about which of a small set of targets a particular message refers to, but how to represent the relevant targets of reference in the first place. Learning to communicate effectively may require discovering a lower-dimensional representation in which the targets of reference vary. For instance, when using sketches to communicate about the identity of complex pieces of music (Healey, Swoboda, Umata, & King, 2007), a particular set of strokes could correspond to any number of properties (pitch, tempo, melody, rhythm, intensity) at any temporal granularity. This is made particularly clear in a classic maze game (Garrod & Anderson, 1987): in order to give effective spatial directions, speakers had well-tuned lexical priors but had to coordinate on what space of *referents* to use (e.g. paths, coordinates, lines, landmarks).

Prior theories have assumed representations of lexical meanings are relatively fixed and the only learning taking place is how one's partner construes a multi-stable percept. For examine, this seems to be what Brennan and Clark (1996) had in mind when they coined the term *conceptual pact*, and (Stolk et al., 2016) have influentially argued that partners in communication construct shared conceptual spaces. Given present data it is not clear how these two sources of uncertainty could be teased apart, though certain kinds of conventions (e.g. proper names or acronyms) seem to rely more on binding new linguistic tokens to meanings than on constructing new conceptualizations. Thus, we expect both levels of coordination are likely to play an important role. Our probabilistic model could be extended to handle additional levels of coordination by placing uncertainty over a hyper-parameter corresponding to the intended feature dimension that must be jointly inferred with the correspondence along that dimension.

Conventions on networks Speakers use different language to talk with different partners in different communities (Auer, 2013). For instance, when a scientist is talking to other scientists about their work, they know they can use efficient technical shorthand that they would avoid when talking to their non-expert friends and family. How and why does such community-specific shorthand form in the first place? And how do we represent which conventions are active for different (novel) partners? While our hierarchical model established mechanisms for convergence *within* a community, a natural follow-up question is how people flexibly switch between different sets of conventions *across* different communities.

Previous work has probed representations of community membership by manipulating the extent to which cultural background is shared between speaker and listener. For example, Isaacs and Clark (1987) paired participants who had either lived in NYC or had never been there for a task referring to landmarks in the city (e.g. "Rockefeller Center"). Within just a few utterances from a novel partner, people could infer whether they were playing with an expert or novice and immediately adjust their language use to be appropriate for this inferred identity. Social information about a partner's group can be so important that even players in artificial-language games react to the restrictions of social anonymity by learning to identify members of their community using distinctive signals (Roberts, 2010).

An important avenue for future work is to use large-scale networked experiments to evaluate the hypothesis that a hierarchical representation of conventions includes not just a partner-specific level and population-wide level but also intermediate community levels. This hypothesis can be formalized by including additional latent representations of community membership into our hierarchical model. That is, in addition to updating our model of a particular *partner* based on immediate feedback, even sparse observations of a partner's language use may license much broader inferences about their lexicon via diagnostic information about their social group or background. If someone's favorite song is an obscure B-side from an 80s hardcore band, you can make fairly strong inferences about what else they like to listen to and how similar they might be to you (Vélez, Bridgers, & Gweon, 2016; Gershman, Pouncy, & Gweon, 2017). Similarly, if someone casually refers to an obscure New York landmark you also recognize, you can safely update your beliefs about their lexicon to include a number of other conventions shared among New Yorkers. Lexica cluster within social groups, so inverting this relationship can yield rapid lexical learning from inferences about social group membership.

Closing thought Language is not some rigid 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 family of related languages, across every repeated interaction with every partner.

References

Abell, C. (2009). Canny resemblance. *Philosophical Review*, *118*(2), 183–223.

Arora, S., Liang, Y., & Ma, T. (2017). A simple but tough-to-beat baseline for sentence embeddings. In *International conference on learning representations (ICLR).*

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

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

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

Barsalou, L. W. (1983). Ad hoc categories. *Memory & cognition*, *11*(3), 211–227.

Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A neural probabilistic language model. *Journal of machine learning research*, *3*(Feb), 1137–1155.

Bergelson, E., & Aslin, R. N. (2017). Nature and origins of the lexicon in 6-mo-olds. *Proceedings of the National Academy of Sciences*, *114*(49), 12916-12921.

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

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, *112*(518), 859–877.

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.

Caldwell, C. A., & Smith, K. (2012). Cultural evolution and perpetuation of arbitrary communicative conventions in experimental microsocieties. *PloS one*, *7*(8), e43807.

Carroll, J. M. (1980). Naming and describing in social communication. *Language and Speech*, *23*(4), 309–322.

Centola, D., & Baronchelli, A. (2015). The spontaneous emergence of conventions: An experimental study of cultural evolution. *Proceedings of the National Academy of Sciences*, *112*(7), 1989–1994.

Chang, A. X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., … others (2015). Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*.

Clark, E. V. (2009). *First language acquisition*. Cambridge University Press.

Clark, E. V., & Clark, H. H. (1979). When nouns surface as verbs. *Language*, 767–811.

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.

Clark, H. H. (1998). Communal lexicons. In K. Malmkjaer & J. Williams (Eds.), *Context in language learning and language understanding* (pp. 63–87). Cambridge: Cambridge University Press.

Clark, H. H., & Gerrig, R. J. (1983). Understanding old words with new meanings. *Journal of verbal learning and verbal behavior*, *22*(5), 591–608.

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

Cohn-Gordon, R., Goodman, N. D., & Potts, C. (2018). Pragmatically Informative Image Captioning with Character-Level Reference. *arXiv preprint arXiv:1804.05417*.

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

Cruse, D. A. (1977). The pragmatics of lexical specificity. *Journal of linguistics*, *13*(02), 153–164.

Cruse, D. A. (1986). *Lexical semantics*. Cambridge: Cambridge University Press.

Danks, J. H. (1970). Encoding of novel figures for communication and memory. *Cognitive Psychology*, *1*(2), 179–191.

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

Degen, J., Hawkins, R. X. D., Graf, C., Kreiss, E., & Goodman, N. D. (2019). When redundancy is rational: A bayesian approach to 'overinformative' referring expressions. *arXiv preprint arXiv:1903.08237*.

Delaney-Busch, N., Morgan, E., Lau, E., & Kuperberg, G. (2017). Comprehenders rationally adapt semantic predictions to the statistics of the local environment: a bayesian model of trial-by-trial n400 amplitudes. In *Proceedings of the 39th annual meeting of the cognitive science society* (Vol. 39).

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 ieee conference on computer vision and pattern recognition* (pp. 248–255).

Dingemanse, M., Blasi, D. E., Lupyan, G., Christiansen, M. H., & Monaghan, P. (2015). Arbitrariness, iconicity, and systematicity in language. *Trends in cognitive sciences*, *19*(10), 603–615.

Fan, J. E., Hawkins, R. X. D., Wu, M., & Goodman, N. D. (2019). Pragmatic inference and visual abstraction enable contextual flexibility during visual communication. *arXiv preprint arXiv:1903.04448*.

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.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998–998.

Frank, M. C., Goodman, N. D., & Tenenbaum, J. B. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, *20*(5), 578–585.

Franke, M., Dablander, F., Schöller, A., Bennett, E., Degen, J., Tessler, M. H., … Goodman, N. D. (2016). What does the crowd believe? a hierarchical approach to estimating subjective beliefs from empirical data. In *Proceedings of the 38th annual meeting of the cognitive science society.*

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

Fraser, B. (2010). Pragmatic competence: The case of hedging. *New approaches to hedging*, *1534*.

Furnas, G. W., Landauer, T. K., Gomez, L. M., & Dumais, S. T. (1987). The vocabulary problem in human-system communication. *Communications of the ACM*, *30*(11), 964–971.

Fussell, S. R., & Krauss, R. M. (1989a). 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.

Fussell, S. R., & Krauss, R. M. (1989b). Understanding friends and strangers: The effects of audience design on message comprehension. *European Journal of Social Psychology*, *19*(6), 509–525.

Fussell, S. R., & Krauss, R. M. (1992). Coordination of knowledge in communication: effects of speakers' assumptions about what others know. *Journal of personality and Social Psychology*, *62*(3), 378.

Futrell, R., Mahowald, K., & Gibson, E. (2015). Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, *112*(33), 10336–10341.

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

Galantucci, B., & Garrod, S. (2011). Experimental semiotics: a review. *Frontiers in human neuroscience*, *5*, 11.

Galashov, A., Jayakumar, S. M., Hasenclever, L., Tirumala, D., Schwarz, J., Desjardins, G., … Heess, N. (2018). Information asymmetry in KL-regularized RL. In *International conference on learning representations (ICLR)*.

Garrod, S., & Anderson, A. (1987). Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition*, *27*(2), 181–218.

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

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.

Garrod, S., Fay, N., Rogers, S., Walker, B., & Swoboda, N. (2010). Can iterated learning explain the emergence of graphical symbols? *Interaction Studies*, *11*(1), 33–50.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (3rd ed.). CRC press Boca Raton, FL.

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

Gerrig, R. J., & Bortfeld, H. (1999). Sense creation in and out of discourse contexts. *Journal of memory and Language*, *41*(4), 457–468.

Gershman, S. J., Pouncy, H. T., & Gweon, H. (2017). Learning the structure of social influence. *Cognitive Science*, *41*, 545–575.

Gibson, E., Futrell, R., Jara-Ettinger, J., Mahowald, K., Bergen, L., Ratnasingam, S., ... Conway, B. R. (2017). Color naming across languages reflects color use. *Proceedings of the National Academy of Sciences*, *114*(40), 10785-10790.

Glucksberg, S., Krauss, R. M., & Weisberg, R. (1966). Referential communication in nursery school children: Method and some preliminary findings. *Journal of experimental child psychology*, *3*(4), 333–342.

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. (1976). *Languages of art: An approach to a theory of symbols.* Hackett publishing.

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

Goodman, N. D., & Lassiter, D. (2014). Probabilistic semantics and pragmatics: Uncertainty in language and thought. In S. Lappin & C. Fox (Eds.), *Handbook of contemporary semantic theory.* Wiley-Blackwell.

Goodman, N. D., & Stuhlmüller, A. (electronic). *The design and implementation of probabilistic programming languages.*

Goodman, N. D., & Stuhlmüller, A. (2013). Knowledge and implicature: Modeling language understanding as social cognition. *Topics in cognitive science*, *5*(1), 173–184.

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

Graf, C., Degen, J., Hawkins, R. X. D., & Goodman, N. D. (2016). Animal, dog, or dalmatian? level of abstraction in nominal referring expressions. In *Proceedings of the 38th annual conference of the Cognitive Science Society.*

Grant, E., Finn, C., Levine, S., Darrell, T., & Griffiths, T. (2018). Recasting Gradient-Based Meta-Learning as Hierarchical Bayes. *arXiv preprint arXiv:1801.08930*.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Syntax and Semantics* (pp. 43–58). New York: Academic Press.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological review*, *114*(2), 211.

Güçlü, U., & van Gerven, M. A. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, *35*(27), 10005–10014.

Hasson, U., Ghazanfar, A. A., Galantucci, B., Garrod, S., & Keysers, C. (2012). Brain-to-brain coupling: a mechanism for creating and sharing a social world. *Trends in cognitive sciences*, *16*(2), 114–121.

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

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

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

Hoffmann, D., Standish, C., García-Diez, M., Pettitt, P., Milton, J., Zilhão, J., ... others (2018). U-th dating of carbonate crusts reveals neandertal origin of iberian cave art. *Science*, *359*(6378), 912–915.

Honnibal, M., & Montani, I. (2019). spaCy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*.

Horton, W. S., & Gerrig, R. J. (2002). Speakers' experiences and audience design: Knowing when and knowing how to adjust utterances to addressees. *Journal of Memory and Language*, *47*(4), 589–606.

Horton, W. S., & Gerrig, R. J. (2005). The impact of memory demands on audience design during language production. *Cognition*, *96*(2), 127–142.

Horton, W. S., & Gerrig, R. J. (2016). Revisiting the Memory-Based Processing Approach to Common Ground. *Topics in Cognitive Science*, *8*(4), 780–795.

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

Hupet, M., Seron, X., & Chantraine, Y. (1991). The effects of the codability and discriminability of the referents on the collaborative referring procedure. *British Journal of Psychology*, *82*(4), 449–462.

Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, *532*(7600), 453–458.

Innes, J. M. (1976). The structure and communication effectiveness of inner and external speech. *British Journal of Social & Clinical Psychology*, *15*(1), 97-99.

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

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

Jones, M. N., Willits, J., Dennis, S., & Jones, M. (2015). Models of semantic memory. *Oxford handbook of mathematical and computational psychology*, 232–254.

Joshi, A., Ghosh, S., Betke, M., Sclaroff, S., & Pfister, H. (2017). Personalizing Gesture Recognition Using Hierarchical Bayesian Neural Networks. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR).*

Jurafsky, D., & Martin, J. H. (2014). *Speech and language processing* (Vol. 3). Pearson London.

Kao, J. T., Wu, J. Y., Bergen, L., & Goodman, N. D. (2014). Nonliteral understanding of number words. *Proceedings of the National Academy of Sciences*, *111*(33), 12002–12007.

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.

King, E., & Sumner, M. (2015). Voice-specific effects in semantic association. In *Proceedings of the 37th annual meeting of the cognitive science society.*

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

Kiros, R., Zhu, Y., Salakhutdinov, R. R., Zemel, R., Urtasun, R., Torralba, A., & Fidler, S. (2015). Skip-thought vectors. In *Advances in neural information processing systems* (pp. 3294–3302).

Kleinschmidt, D. F. (2019). Structure in talker variability: How much is there and how much can it help? *Language, cognition and neuroscience*, *34*(1), 43–68.

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.

Koolen, R., Gatt, A., Goudbeek, M., & Krahmer, E. (2011). Factors causing overspecification in definite descriptions. *Journal of Pragmatics*, *43*(13), 3231–3250.

Krauss, R. M., & Bricker, P. D. (1967). Effects of transmission delay and access delay on the efficiency of verbal communication. *The Journal of the Acoustical Society of America*, *41*(2), 286–292.

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., & Glucksberg, S. (1977). Social and nonsocial speech. *Scientific American*, *236*(2), 100–105.

Krauss, R. M., Vivekananthan, P., & Weinheimer, S. (1968). Inner speech and" external speech": Characteristics and communication effectiveness of socially and nonsocially encoded messages. *Journal of Personality and Social Psychology*, *9*(4), 295.

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.

Krauss, R. M., & Weinheimer, S. (1967). Effect of referent similarity and communication mode on verbal encoding. *Journal of Verbal Learning and Verbal Behavior*, *6*(3), 359–363.

Kubilius, J., Bracci, S., & de Beeck, H. P. O. (2016). Deep neural networks as a computational model for human shape sensitivity. *PLoS Computational Biology*, *12*(4), e1004896.

Kuhlen, A. K., & Brennan, S. E. (2013). Language in dialogue: when confederates might be hazardous to your data. *Psychonomic bulletin & review*, *20*(1), 54–72.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40*.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, *104*(2), 211.

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

Lassiter, D., & Goodman, N. D. (2015). Adjectival vagueness in a Bayesian model of interpretation. *Synthese*, 1–36.

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

Liesefeld, H. R., & Janczyk, M. (2019). Combining speed and accuracy to control for speed-accuracy trade-offs. *Behavior Research Methods*, *51*(1), 40–60.

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).

Maaten, L. v. d., & Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, *9*(Nov), 2579–2605.

Medlock, B., & Briscoe, T. (2007). Weakly supervised learning for hedge classification in scientific literature. In *Acl* (Vol. 2007, pp. 992–999).

Mervis, C. B., & Rosch, E. (1981). Categorization of natural objects. *Annual Review of Psychology*, *32*(1), 89-115.

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.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111–3119).

Misyak, J., Noguchi, T., & Chater, N. (2016). Instantaneous conventions: The emergence of flexible communicative signals. *Psychological Science*.

Monroe, W., Hawkins, R. X. D., Goodman, N. D., & Potts, C. (2017). Colors in Context: A Pragmatic Neural Model for Grounded Language Understanding. *arXiv preprint arXiv:1703.10186*.

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

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

Osgood, C. E. (1952). The nature and measurement of meaning. *Psychological bulletin*, *49*(3), 197.

Partee, B. (1995). Lexical semantics and compositionality. In *An invitation to cognitive science, part i: Language*. Cambridge, MA: MIT Press.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (emnlp)* (pp. 1532–1543).

Perlman, M., Dale, R., & Lupyan, G. (2015). Iconicity can ground the creation of vocal symbols. *Royal Society open science*, *2*(8), 150152.

Peterson, J. C., Abbott, J. T., & Griffiths, T. L. (2018). Evaluating (and improving) the correspondence between deep neural networks and human representations. *Cognitive Science*, *42*(8), 2648–2669.

Pinker, S. (1995). *The language instinct: The new science of language and mind*. Penguin UK.

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.

Potts, C., & Levy, R. (2015). Negotiating lexical uncertainty and speaker expertise with disjunction. In *Proceedings of the 41st annual meeting of the berkeley linguistics society* (Vol. 41).

Prelec, D. (2004). A bayesian truth serum for subjective data. *science*, *306*(5695), 462–466.

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

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.

Ritchie, D., Horsfall, P., & Goodman, N. D. (2016). Deep amortized inference for probabilistic programs. *arXiv:1610.05735*.

Roberts, G. (2010). An experimental study of social selection and frequency of interaction in linguistic diversity. *Interaction Studies*, *11*(1), 138–159.

Roberts, G., & Galantucci, B. (2012). The emergence of duality of patterning: Insights from the laboratory. *Language and cognition*, *4*(4), 297–318.

Roberts, G., Lewandowski, J., & Galantucci, B. (2015). How communication changes when we cannot mime the world: Experimental evidence for the effect of iconicity on combinatoriality. *Cognition*, *141*, 52–66.

Robertson, S. (2004). Understanding inverse document frequency: on theoretical arguments for idf. *Journal of documentation*, *60*(5), 503–520.

Robins, A. (1995). Catastrophic forgetting, rehearsal and pseudorehearsal. *Connection Science*, *7*(2), 123–146.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information processing & management*, *24*(5), 513–523.

Schober, M. F., & Clark, H. H. (1989). Understanding by addressees and overhearers. *Cognitive Psychology*, *21*(2), 211–232.

Setoh, P., Scott, R. M., & Baillargeon, R. (2016). Two-and-a-half-year-olds succeed at a traditional false-belief task with reduced processing demands. *Proceedings of the National Academy of Sciences*, *113*(47), 13360-13365.

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).

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, *28*(1), 11–21.

Spike, M., Stadler, K., Kirby, S., & Smith, K. (2017). Minimal requirements for the emergence of learned signaling. *Cognitive science*, *41*(3), 623–658.

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

Steels, L. (2003). Evolving grounded communication for robots. *Trends in Cognitive Sciences*, *7*(7), 308–312.

Steels, L. (2015). *The talking heads experiment: Origins of words and meanings* (Vol. 1). Language Science Press.

Steels, L., & Belpaeme, T. (2005). Coordinating perceptually grounded categories through language: A case study for colour. *Behavioral and brain sciences*, *28*(4), 469–488.

Stolk, A., Verhagen, L., & Toni, I. (2016). Conceptual alignment: How brains achieve mutual understanding. *Trends in cognitive sciences*, *20*(3), 180–191.

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.

Tesink, C. M., Petersson, K. M., Van Berkum, J. J., Van den Brink, D., Buitelaar, J. K., & Hagoort, P. (2009). Unification of speaker and meaning in language comprehension: An fmri study. *Journal of Cognitive Neuroscience*, *21*(11), 2085–2099.

Tessler, M. H., & Goodman, N. D. (2019). The language of generalization. *Psychological review*, *126*(3), 395.

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

Van Berkum, J. J., Van den Brink, D., Tesink, C. M., Kos, M., & Hagoort, P. (2008). The neural integration of speaker and message. *Journal of cognitive neuroscience*, *20*(4), 580–591.

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

Vedantam, R., Bengio, S., Murphy, K., Parikh, D., & Chechik, G. (2017). Context-aware Captions from Context-agnostic Supervision. *arXiv preprint arXiv:1701.02870*.

Vélez, N., Bridgers, S., & Gweon, H. (2016). Not all overlaps are equal: Social affiliation and rare overlaps of preferences. In *Proceedings of the 38th annual conference of the cognitive science society. austin, tx: Cognitive science society.*

Verhoef, T., Kirby, S., & de Boer, B. (2016). Iconicity and the emergence of combinatorial structure in language. *Cognitive science*, *40*(8), 1969–1994.

Verhoef, T., Roberts, S. G., & Dingemanse, M. (2015). Emergence of systematic iconicity: Transmission, interaction and analogy. In D. C. Noelle et al. (Eds.), *Proceedings of the 37th annual conference of the cognitive science society.*

Verhoef, T., Walker, E., & Marghetis, T. (2016). Cognitive biases and social coordination in the emergence of temporal language. In *The 38th annual meeting of the cognitive science society (cogsci 2016)* (pp. 2615–2620).

Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and Tell: A Neural Image Caption Generator. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition* (pp. 3156–3164).

Weber, R. A., & Camerer, C. F. (2003). Cultural conflict and merger failure: An experimental approach. *Management Science*, *49*(4), 400–415.

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.

Xu, F., & Tenenbaum, J. B. (2007). Word learning as bayesian inference. *Psychological Review*, *114*(2), 245.

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 201403112.

Yoon, S. O., & Brown-Schmidt, S. (2014). Adjusting conceptual pacts in three-party conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(4), 919.

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

Yu, D., Yao, K., Su, H., Li, G., & Seide, F. (2013). Kl-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition. In *2013 ieee international conference on acoustics, speech and signal processing* (pp. 7893–7897).

This thesis was typeset using LaTeX, originally developed by Leslie Lamport and based on Donald Knuth's TeX. The body text is set in 11 point Egenolff-Berner Garamond, a revival of Claude Garamont's humanist typeface. The above illustration, *Science Experiment 02*, was created by Ben Schlitter and released under CC BY-NC-ND 3.0. A template that can be used to format a PhD dissertation with this look & feel has been released under the permissive AGPL license, and can be found online at github.com/suchow/Dissertate or from its lead author, Jordan Suchow, at suchow@post.harvard.edu.