Co-evolution of language and mindreading: A computational exploration

Marieke S. Woensdregt

Submitted in fulfilment of the degree of Doctor of Philosophy to
School of Philosophy, Psychology and Language Sciences
University of Edinburgh
2019

Declaration

I declare that the thesis has been composed by myself and that the work has not been submitted for any other degree or professional qualification. I confirm that the work submitted is my own, except where work which has formed part of jointly-authored publications has been included. My contribution and those of the other author to this work have been explicitly indicated below. I confirm that appropriate credit has been given within this thesis where reference has been made to the work of others. The work presented in Chapter 2 was previously published in the Oxford Research Encyclopedia of Linguistics as Pragmatics and Language Evolution by Marieke Woensdregt (author of this thesis and declaration) and Kenny Smith (PhD supervisor). This encyclopedia chapter was conceived by both of the authors. I carried out the writing and the literature research. The model presented in Chapter 3 was previously published in the Proceedings of the 38th Annual Meeting of the Cognitive Science Society as Modelling the co-development of word learning and perspective-taking by Marieke Woensdregt (author of this thesis and declaration), Simon Kirby, Chris Cummins and Kenny Smith (all three supervisors of this PhD thesis). The work presented in that proceedings paper is a precursor to the work presented in Chapter 3. The modelling work included in the paper was conceived by all four authors together, and I carried out the writing work. A pdf of this proceedings paper is included in appendix F. All the code that was used to run the simulations presented in this thesis is freely available on my GitHub page: github.com/marieke-woensdregt.

(*Marieke S. Woensdregt*)

Lay Summary

One thing that's special about humans is the fact that we have language, which far exceeds the communication systems of other animals in its complexity and open-endedness. Another thing that's special about humans is the fact that we are very skilled at thinking about what's going on in other people's minds. We are able to figure out what other people feel, know and believe, and we understand that another person's feelings, knowledge and beliefs can be different from our own. This ability is known as mindreading, and we use it all the time to interpret and predict others' behaviour. Although other social animals have some mindreading abilities, these are limited compared to those of humans.

Language and mindreading are not independent from each other. To successfully get a message across using language, we need to think about what's going on in the mind of our conversation partner. And the other way around, language helps us to find out and think about what's going on in other people's minds, and to pass on this knowledge to our children (and thus over evolutionary time). Because of this interdependence between language and mindreading, scientists have speculated that the two skills have played a role in enabling each other's evolution (that is, that they have *co-evolved*). However, this has been hard to establish, because we only have very indirect evidence of how our early ancestors thought and communicated. In this thesis, I use a model in which I simulate a population of artificial 'agents' (a bit like bodyless robots living inside a computer) and their evolution, in order to explore under what circumstances a co-evolutionary dynamic between language and mindreading could have come about.

In my model, I gave agents not just the ability to learn a language and communicate with it, but also a private 'perspective on the world', which influences what they talk about in a given context. Agents' ability to learn about each other's perspectives instantiates a simple form of mindreading. The simulation results of this model show that these agents' language and mindreading *co-develop*. That is, the agents cannot

learn a language properly if they cannot learn about a speaker's perspective, and they cannot learn about the speaker's perspective properly if the speaker's language is meaningless (for instance, if it uses only one word to refer to everything). This brings us to an evolutionary question: if this co-developmental process depends on the language being meaningful, how could a population of these agents develop such a language from scratch?

I address this question using an evolutionary version of the model in which languages are passed on over generations in a population of agents. Because we are now looking at evolution, we can explore different *selection pressures*; that is, measures of agents' success which determine which agents get to pass on their language to the next generation. Under a selection pressure for communication, agents who are more successful at communicating are more likely to be chosen to teach their language to learners of the next generation, while under a selection pressure for mindreading, agents who are better mindreaders are more likely to pass on their language. The simulation results of this evolutionary model show that under selection for communication, agents can evolve a meaningful language over generations, even if they start out with no language whatsoever. This evolution of a meaningful language leads to the population being more successful not just at communicating, but also at mindreading. This is because sharing a language with others provides agents with information about those others' perspectives. The other way around, under selection for mindreading, a meaningful language also evolves. This again leads not just to successful mindreading, but also to successful communication.

Taken together, these results suggest that co-evolution between language and mindreading can get off the ground even if there is only a pressure for one of these two skills to evolve. In real human evolution, such a pressure may have come from the need to cooperate and coordinate socially, which is likely to have arisen when our early ancestors started hunting big game during the Ice Age.

Abstract

Language relies on mindreading: in order to use it successfully we need to be able to entertain and recognise communicative intentions. Mindreading abilities in turn profit from language, as language provides a means for expressing mental states explicitly, and for transmitting our knowledge of mental states to others. Given this interdependence, it has been hypothesised that language and mindreading have co-evolved. In this thesis I formalise the relationship between language and mindreading in a computational model, in order to explore under what circumstances a co-evolutionary dynamic between the two skills could have gotten off the ground.

In Chapter 3 I present an agent-based model which combines referential signalling with perspective-taking, where perspective-taking instantiates a very simple form of mindreading. In this model, agents' communicative behaviour is probabilistically determined by an interplay between their language and their perspective on the world. The *literal* variant of these agents (explored in Chapters 3 and 4) consists of speakers who produce utterances purely based on their own language and perspective, and listeners who interpret these utterances using what they've learned about the speaker's perspective through interaction. The *pragmatic* variant of these agents in contrast (explored in Chapters 5 and 6) consists of speakers who optimise their utterances by maximising the probability that the listener will interpret them correctly (assuming the listener shares their perspective), and listeners who interpret these utterances by reasoning about such a speaker, again using what they've learned about the speaker's perspective through interaction. Learning is not straightforward however, because agents' languages and perspectives are private (i.e. not directly observable to other agents). Instead, the Bayesian learners in this model only get to observe a speaker's utterances in context, from which they have to simultaneously infer the speaker's language and perspective. Simulation results show that learners can overcome this joint inference problem by bootstrapping one from the other, but that the success of this process depends on how

informative the speaker's language is.

This leads to an evolutionary question: If the co-development between language-learning and perspective-learning relies on agents being exposed to an informative language, how could a population of such agents evolve an informative language from scratch? I address this question with an iterated learning version of the model described above, combined with different selection pressures. Simulation results with literal agents (presented in Chapter 4) show that an informative language emerges not just if the population is subjected to a selection pressure for communication, but also under selection for accurate perspective-inference. Under both pressures, the emergence of an informative language leads not just to more successful communication, but also to more successful perspective-inference. This is because sharing an informative language with others provides agents with information about those others' perspectives (note that agents' innate ability to learn about others' perspectives does not change over generations). Simulation results with pragmatic agents (presented in Chapter 5) show the same co-evolutionary dynamics as literal agents, with the difference that they can achieve equally high levels of success at communicating and inferring perspectives with much more ambiguous languages, because they can compensate for suboptimal languages using their pragmatic ability. Finally, in Chapter 6 I explore under what circumstances such pragmatic agents could have evolved; that is, under what circumstances being a pragmatic communicator provides an evolutionary advantage over being a literal communicator.

Taken together, the model results presented in this thesis suggest firstly that co-evolution between language and mindreading could have gotten off the ground under any circumstances which created a need for either improved communication *or* improved insight into others' minds. Secondly, the results suggest that such a co-evolutionary dynamic could have been driven largely by cultural evolution; where mindreading improves by virtue of evolving a language.

Acknowledgements

First of all, I want to thank my supervisors, Simon, Kenny and Chris, for sharing their skills, knowledge and wisdom so generously, and for always showing faith in me and my PhD project. I want to thank Jenny for doing exactly the same, despite not officially being one of my supervisors. In particular I want to thank Simon for seeing the bigger picture when I didn't, Kenny for seeing the possibilities where I didn't, and Chris for always keeping an eye on the detail.

I feel very lucky to have been part of the LEC/CLE throughout my PhD years (and before). I have learned so much from all of you about how to be a good researcher and academic, *and* about how to continue being a fully-rounded human being in the process. In roughly chronological order, thank you Monica, Olga, Big Marieke, Chrissy, Molly, Isabelle, Stella, Alex, Jia, Mora, and Matt for each contributing in your own way to making the CLE such a wonderful group to work in.

Special thanks go to all the CLE PhD students, past and present, who have been such a fun, warm, stimulating and inspiring peer group throughout the years. Thank you Cat, Bill, Vanessa, Mark, Matt, James, Alan, Kevin, Yasamin, Jasmeen, Carmen, Jon, Cathleen, Fiona, Fausto, Ash, Jonas, Andres, Svenja, Tamar and Fang. With special thanks to Bill and Matt for being my 'big brothers' in modelling, to Cathleen and Fausto for being my topic (and singing!) buddies, to Jasmeen for all the yoga sessions, and to Yasamin for being the cornerstone of office life for so many years.

Speaking of cornerstones, special thanks also to E, my one and only writing pal, for all the days we spent sitting across from each other in DSB 3.01 and other places, in silent yet joint writing effort. Without those days this thesis would not have been here. Thanks also to everyone else in the DSB 1.15 PhD office for making it such a pleasant place to work in. And thank you Katie Keltie and Toni Noble, for making sure all this runs smoothly.

This thesis has of course also greatly benefited from conversations with fellow re-

searchers outside of the Centre for Language Evolution. In particular I would like to thank Bill Thompson, Thomas Brochhagen, Robert Hawkins, Richard Moore, Michael Franke, Sean Roberts, Jennie Pyers, and Ann Senghas, for thought-provoking and fruitful discussions.

I also want to thank my parents Hans and Margo, my brothers, Tim and Jinko, and of course Marieke J. and Olga, for showing me both curiosity and persistence, and supporting me from the very beginning. Speaking of being there from the beginning, I want to thank Djoeke, Iduna, Linus, Alexandra, Kim and Sanne for always being there to take me out of my bubble, always providing me with a place to stay, and simply being my homies, no matter where I am. And special thanks to Djoeke for helping me with my lay summary!

I equally want to thank all my friends in Edinburgh, who together have made this city into a home: Gabby (with special thanks for giving me access to Brody — bundle of love!), Jasmeen, Nikita (and Murdo!), Anna, Cleo, Alessio, Alessia, Berta, Lucía, Ricardo (and Noelia!), Panos, Yasamin, E, Alex, Mora, and Emma.

Last and foremost, I want to thank Carmen. Without you by my side my PhD years would have been a scarier, harder, and — let's face it — duller time. I cannot thank you enough for all the support you've given me over the years, and for making my life so much richer.

Contents

Chapter 1

Introduction

The hypothesis that language and mindreading (also known as 'theory of mind') have co-evolved, has been put forward by several different theorists of human evolution (see e.g. Malle, 2002; Moore, 2016a; Whiten and Erdal, 2012). This hypothesis is captured concisely in the following quote by Whiten and Erdal (2012):

> Mindreading has been argued to underwrite the intentionality of human language, in which utterances are delivered with the intent that others will take certain meanings from them. In turn, terminology and talk about what is in or on our minds is embodied in language.
>
> — Whiten and Erdal (2012, p. 2126)

This quote brings together two ideas. Firstly, language as we see it in humans today is made possible by our ability to entertain and recognise communicative intentions (which is important for both language learning and language use). These are in effect intentions to manipulate the mental states of others, and therefore require certain mindreading abilities. Mindreading here refers to the process of ascribing mental states (thoughts, beliefs, desires, etc.) to oneself and others. Secondly, language in turn provides us with a means of expressing our mental states explicitly, and with terminology that allows us to convey our own understanding of mental states to others (e.g. mental state terms such as "think" and "know", and sentential complement constructions such as "Ella believes that x"). Thus, language itself may further improve mindreading abilities because it both provides clear data to learn from, and enables cultural transmission of our understanding of minds to younger members of the population. These theoretical considerations suggest that the evolution of language and mindreading in

humans has benefited from a two-way positive feedback loop, which may have been driven by cultural, rather than biological, evolution.

This co-evolutionary scenario has been fleshed out in some detail by Malle (2002) and Moore (2016a), but is difficult, if not impossible, to test empirically, because hominins in earlier stages of these evolutionary processes are no longer alive. And although some have left artefacts and other traces from which we can make inferences about their social and inner life, their cognition has not fossilised. The most informative test-case we can observe today may be that of emerging sign languages (such as Nicaraguan Sign Language), which have started their process of evolution relatively recently, allowing us to study both the very first and the later signers of these languages (Pyers and Senghas, 2009). However, these languages have still emerged within the wider context of a highly encultured society, which is not comparable to when language first emerged. This thesis therefore chooses a different route: that of computational modelling. Specifically, the aim of this thesis is to formalise the preliminaries of the co-evolutionary hypothesis outlined above in an agent-based model, in order to explore under what circumstances such a co-evolutionary dynamic between language and mindreading could have gotten off the ground.

In Chapter 2, I outline the theoretical background for the hypothesis that language and mindreading have co-evolved. This chapter starts with an exposition of the claim that language use as we see it in humans today requires rather sophisticated mindreading. It then reviews the empirical evidence of such mindreading abilities in our closest living relatives: the nonhuman great apes. This leads to the conclusion that although nonhuman primates have at least some of the precursors to the full-blown mindreading and pragmatic capacity we find in humans, there seem to be limitations on their abilities which may restrict the extent to which they can entertain and recognise communicative intentions. Thus, after diverging from our last common ancestor with the nonhuman primates, hominins must have undergone further evolution of our mindreading and pragmatic abilities. Chapter 2 then continues with a discussion of the role that biological evolution and cultural evolution may have played in this process.

The final sections of Chapter 2 (as well as parts of Chapter 3) review compelling evidence in favour of the hypothesis that the explicit mindreading abilities we find in humans today are the product of cultural rather than biological evolution (Heyes and Frith, 2014; Heyes, 2018). That is, that human mindreading skills have reached their current level of sophistication through a process of cultural transmission, where new

generations build on the understanding of minds that was passed on to them by the previous generation. This process could have resulted in a complex suite of mindreading knowledge and skills being accumulated over evolutionary time. This hypothesis is of specific importance to the argument made in this thesis, because if our mindreading skills have been accumulated over generations through cultural transmission, language is likely to have facilitated that process. After all, language is a tool that greatly enhances the range of things that can be culturally transmitted, and the precision with which this transmission can happen (Heyes and Frith, 2014).

Once this theoretical groundwork has been laid, Chapter 3 focuses on development rather than evolution, and starts with a review of how language and mindreading interact during development. This review reveals evidence of co-development. On the one hand, typically developing children use their ability to infer communicative intentions to help them learn the meaning of words. Deaf children who grow up without being exposed to an existing sign language and community of signers suffer from delays in their mindreading development, which in some extreme cases may never be caught up on. Thus, on the other hand, mindreading abilities play a role in language development, but exposure to language also plays a role in mindreading development.

Based on this evidence of co-development, Chapter 3 introduces the agent-based model of language-learning which forms the core model that the rest of this thesis builds on. This model integrates the role of mental states in linguistic communication in a simple way based on two premises: *what* an agent refers to in a given context depends (probabilistically) on their perspective on the world, *how* the agent refers to their chosen referent depends on their lexicon. Because this model of how utterances are produced in contexts involves a subjective, unobservable attribute of the speaker (their perspective), a learner who wants to acquire the speaker's lexicon is forced to simultaneously learn about the speaker's perspective. However, the only data the learner gets to observe that might inform them about the speaker's perspective are the speaker's utterances in context. Thus, the learner is also forced to use their developing knowledge of the lexicon in order to bootstrap their perspective-learning. Given enough data, Bayesian learners can solve this joint inference task, as long as two conditions are met. Firstly, the learner needs to be able to represent the speaker's perspective, and secondly, the speaker's lexicon needs to be at least somewhat informative.

The latter condition leads to the question: under what circumstances will a population of agents who develop in this way evolve an informative lexicon from scratch

(i.e. if they all start out with a completely ambiguous lexicon). This question is the focus of Chapter 4, where the developmental model described above is embedded in an iterated learning model — in which lexicons are passed on over generations through learning from the behaviour of the previous generation. Simulation results of this model show that when such iterated learning is combined with either a selection pressure for successful communication, or a selection pressure for successful perspective-inference, populations evolve informative lexicons. Because both lexicon-learning and perspective-inference benefit from having an informative lexicon, both skills improve under either selection pressure. That is, a pressure for communication leads to the evolution of informative lexicons, which in turn leads to more accurate perspective-inference, and a pressure on perspective-inference equally leads to the evolution of informative lexicons (because this is the only way in which agents can accumulate evidence about each others' perspectives), and thereby in turn increases the population's communicative success.

In Chapter 5, the model of communication is expanded by adding a layer of pragmatic reasoning on top, for both speaker and listener. Simulations of iterated learning with this model show similar results in the sense that a pressure for communication drives improvement in perspective-inference and vice versa, but populations of pragmatic agents achieve this effect with more ambiguous lexicons. These populations can compensate for such ambiguity in the lexicon using their pragmatic ability, and are thus under less pressure to evolve completely unambiguous lexicons.

Chapter 6 explores the extent to which such pragmatic agents have an evolutionary advantage over the 'literal' agents explored in Chapter 4. In order to answer this question, I use a model of gene-culture co-evolution combined with an invasibility analysis. The results of this analysis show that pragmatic agents have an evolutionary advantage over literal agents under both a pressure for communication and a pressure for perspective-inference. However, surprisingly, an increase in the amount of pragmatic agents in a population does not lead to an overall increase in the populations' success at these two skills. Instead, as the amount of pragmatic agents in the population increases, the pressure for maintaining the fully informative, unambiguous lexicons that their literal predecessors had built decreases, and ambiguity seeps into the populations' lexicons. As this ambiguity builds up, the evolutionary advantage of pragmatic agents over literal agents increases (because they are more resistant to ambiguity), leading to yet more pragmatic agents and yet more ambiguity in their lexicons. These results

suggest that a build-up of pragmatic competence in combination with the cultural evolution of languages can lead populations into a 'cultural trap' (Lachlan and Slater, 1999), in the sense that languages evolve in a direction where they require pragmatic competence in order to be used successfully, and there is no turning back.

Finally, Chapter 7 pulls all these results together and discusses their implications for theories about how language and mindreading have evolved in humans.

Chapter 2

Pragmatics and Language Evolution

Marieke Woensdregt and Kenny Smith

The following is a peer-reviewed chapter which was published online as part of the Oxford Research Encyclopedia of Linguistics in March 2017
(DOI: 10.1093/acrefore/9780199384655.013.321). Included below is a preprint of the chapter which was uploaded to the PsyArXiv server on 24 December 2018 (DOI: 10.31234/osf.io/5sqfj), with permission of Oxford University Press. This chapter was conceived together with co-author Kenny Smith, and written by me, Marieke Woensdregt. The citations may be looked up in the reference list included in the chapter itself, on pages 46-53 or in the references list at the end of this thesis. The footnotes may be looked up at the end of the chapter on page 54. Note also that the empirical evidence for co-development of language and mindreading which is briefly discussed in section 7 ("Have Language and Theory of Mind Co-Evolved") of this chapter, is reviewed in more detail in sections 3.1 and 3.2 of Chapter 3.

Pragmatics and Language Evolution

Marieke Woensdregt and Kenny Smith

Summary

Pragmatics is the branch of linguistics that deals with language use in context. It looks at the meaning linguistic utterances can have beyond their literal meaning (implicature), and also at presupposition and turn taking in conversation. Thus, pragmatics lies on the interface between language and social cognition.

From the point of view of both speaker and listener, doing pragmatics requires reasoning about the minds of others. For instance, a speaker has to think about what knowledge they share with the listener to choose what information to explicitly encode in their utterance and what to leave implicit. A listener has to make inferences about what the speaker meant based on the context, their knowledge about the speaker, and their knowledge of general conventions in language use. This ability to reason about the minds of others (usually referred to as "mindreading" or "theory of mind") is a cognitive capacity that is uniquely developed in humans compared to other animals.

This article will review what we know about how pragmatics (and the underlying ability to make inferences about the minds of others) has evolved. Biological evolution and cultural evolution are the two main processes that can lead to the development of a complex behavior over generations, and we can explore to what extent they account for what we know about pragmatics.

In biological evolution, changes happen as a result of natural selection on genetically transmitted traits. In cultural evolution on the other hand, selection happens on skills that are transmitted through social learning. Many hypotheses have been put forward about the role that natural selection may have played in the evolution of social and communicative skills in humans (for example, as a result of changes in food sources, foraging strategy, or group size). The role of social learning and cumulative culture, however, has been often overlooked. This omission is particularly striking in the case of pragmatics, as language itself is a prime example of a culturally transmitted skill, and there is solid evidence that the pragmatic capacities that are so central to language use may themselves be partially shaped by social learning.

In light of empirical findings from comparative, developmental, and experimental research, we can consider the potential contributions of both biological and cultural evolutionary mechanisms to the evolution of pragmatics. The dynamics of types of evolutionary processes can also be explored using experiments and computational models.

Keywords

pragmatics, ostensive-inferential communication, primate communication, theory of mind, biological evolution, cultural evolution, co-evolution

1. Pragmatics and Mindreading

Being a competent language user does not just involve having access to a vocabulary and a grammar that are shared with others. It also involves knowing how to deploy those linguistic

tools to achieve your communicative goals. This requires you to keep track of what your interlocutor knows and doesn't know, how their view on the world differs from your own, and what is appropriate to say in a given situation. In other words, we take into account the context in which communication occurs and exploit its affordances to get our message across. The word *context* here refers not only to the situation and physical surroundings, but also to the mental context of the communicators, that is, what they can see at this moment and also what they are likely to know or be interested in. The field of pragmatics is concerned with how we use such context when producing and interpreting linguistic utterances.

Deploying communicative signals flexibly, depending on context, is not unique to human communication. For instance, captive chimpanzees have been found to use modality flexibly based on the orientation and attentional state of their audience. Leavens, Russell, and Hawkins (2010) found that if a human experimenter was facing a chimpanzee, the latter would use gestures to request a specific food item; if, however, the experimenter was facing away, they used vocalizations to attract the experimenter's attention first. Chimpanzees in the wild have also been observed to adapt their signalling behavior according to the composition of their audience. When attacked by a group member, chimpanzees will normally scream in response, and the acoustic properties of their scream reflect the severity of the aggression, a correlation that nearby group members use to determine whether they should intervene in the fight or not. However, victims will also exaggerate the length and frequency of their screams in response to mild aggression if they know that there's a high-ranking group member around who will be likely to help (Slocombe & Zuberbühler, 2007).

Thus, the ability to flexibly adapt signal choice and the way in which signals are used based on the context is a skill we share at least with our closest living relatives, and therefore

presumably reflects cognitive capacities already present in our last common ancestor with chimpanzees. However, pragmatics in language involves a cognitive capacity that is more restricted in its distribution and possibly unique to humans: the ability to adapt signal use based on knowledge or inferences of what goes on *inside the minds* of others. This is the part of pragmatics that is concerned with implicature and inference, which make use not of observable features of the physical or linguistic context, but of unobservable mental states.

An example of such implicature is the fact that the sentence "Ella got the car to stop" brings with it the implication that Ella did not simply hit the brakes, but got the car to stop in some more unusual fashion. This implicature is known as a *manner implicature,* as it arises from Grice's "Maxim of Manner," which states that speakers should "be perspicuous" and "be brief (avoid unnecessary prolixity)" (Grice, 1975, p. 308). The simplest way to say that Ella stopped the car by hitting the brakes would be to say "Ella stopped the car." Since the speaker said something more elaborate ("Ella got the car to stop"), and thus would be violating the manner maxim if their intended meaning was that Ella hit the brakes, the hearer can infer that the speaker intended to communicate a more complex meaning: that Ella got the car to stop in an unusual way.

Understanding and using such implicature is qualitatively different from adapting one's signal use to the composition of one's audience or their attentional state, because it requires both speaker and hearer to reason about each other's mental states, which are not directly observable but have to be inferred[1]. Several researchers have argued that even the simplest exchanges in human language require several levels of embedded reasoning about mental states (Scott-Phillips, 2015a; Sperber & Wilson, 1995), and that this is what makes human language special when compared to the communication systems of other animals (Scott-Phillips, 2015b).

Although this analysis of what everyday language use consists of is a matter of debate (see e.g., Moore, 2014, 2016a, 2016c), it is not contested that natural linguistic exchanges between humans *can* involve complex inference-making, and that this requires the ability to reason about the content of other's minds (e.g., Moore, 2016c). This ability is often referred to as *theory of mind*, *mindreading*, *metapsychology,* or *mentalizing* (see e.g., Baron-Cohen, Leslie, & Frith, 1985)—in this article we will use the terms *theory of mind* (abbreviated ToM) and *mindreading* interchangeably, simply because these are the most commonly used.

Humans are more proficient mindreaders than any other species. How has this pragmatic competence evolved? Is it a biological adaptation, and if so, what selection pressure has it evolved in response to? Or is it a product of cultural evolution, where skills are transmitted from generation to generation through social learning, accumulating improvements as they go? Or have culture and biology worked together to produce this unique capacity? Have the socio-cognitive abilities that underlie pragmatic competence in humans evolved for the purpose of language, or did they initially evolve for other purposes? Or have language and social cognition co-evolved, the one skill building on the other?

To answer these questions, we will start by providing an analysis of what human pragmatic competence consists of (section 2, "What Is Pragmatic Competence?"), followed by a breakdown of the psychological mechanisms involved (section 3, "Psychological Mechanisms Underlying Pragmatic Competence"). We will then go on to explore to what extent these psychological mechanisms are shared between humans and other primates (section 4, "Pragmatic Competence in Great Apes") to identify which parts of pragmatic competence have evolved exclusively in the *Homo* lineage. Subsequently, we will turn to theories of the evolution of the human-specific components of pragmatic competence. We will first review explanations involving biological

adaptation (section 5, “The Biological Evolution of Human Pragmatic Skills”), followed by explanations drawing on cultural evolution (section 6, “The Cultural Evolution of Human Pragmatic Skills”). Finally, we will discuss the possibility that the socio-cognitive skills underlying pragmatics have co-evolved with language itself (i.e., the conventional code) (section 7, “Have Language and Theory of Mind Co-Evolved?”).

2. What Is Pragmatic Competence?

Pragmatic competence is what allows an individual to look beyond the literal meaning of an utterance to determine the *speaker meaning*. Where literal meaning refers to the semantic concepts that are associated with the words and structure of a sentence, speaker meaning refers to the goal that the speaker has when they produce that sentence. This can be a goal to inform (“the entrance is on the other side of the building”); a request (“could you open the window?”); or general social bonding (“so sunny today!”).

The ability to infer a speaker’s intention behind an utterance obviously comes into play when interpreting deliberately non-literal language use, such as metaphors or sarcasm. But it is also necessary for interpreting a straightforward utterance such as “I’m tired.” Depending on the context, this could mean anything from “Let’s have a coffee break,” to “I don’t feel like talking about it,” to “I’m thinking of quitting my job,” and so on. Thanks to this flexibility in use and interpretation, there may be an infinite set of potential speaker meanings for any given utterance in human language. This phenomenon is known as linguistic *underdeterminacy* (Carston, 2002, pp. 19-30). A hearer can resolve part of this underdeterminacy based on the context and the

preceding conversation, the remainder must be disambiguated based on knowledge and inferences about the speaker's mind.

The phenomenon of linguistic underdeterminacy illustrates that to analyze human communication we must go beyond what is known as the *code model* of communication (Shannon, 1948). In the code model, communication consists of a signaler encoding a message into a signal and a receiver decoding it to uncover the message (often by doing the inverse of the encoding operation). Communication systems that are sufficiently described by this model, sometimes known as *natural codes*, simply consist of pairs of associations, where the signaler has associations between states of the world and signals, and the receiver has associations between signals and responses. Many of the communication systems we find in nonhuman animals can be analyzed in this way (Wharton, 2003). If the encoding and decoding operations in a natural code are properly tuned and there is no noise in transmission, the message that goes in at one end should be the same as what comes out the other. What this model cannot account for is the underdeterminacy of human language—where the same signal can have many different interpretations depending on the situational context, the linguistic context, the manner of delivery, etc. A natural code is based on associations between signals and relevant phenomena in the world. A conventional code (like language) on the other hand, is made possible by associations between signals and inferred speaker meanings (see Wheeler's commentary on Scott-Phillips, 2015b, p.74)

The ability to make inferences about speaker intention is, therefore, an essential part of human language and our pragmatic competence. This requires theory of mind both on the part of the hearer and the part of the speaker. Building on an initial proposal by Grice (1957), Sperber and Wilson (1995) argue that any linguistic utterance contains the following two intentions:

"*Informative intention:* to inform the audience of something;

Communicative intention: to inform the audience of one's informative intention."

Sperber and Wilson (1995, p. 29)

The informative intention contains *what* the speaker wants to communicate, and the communicative intention contains *that* they want to communicate. Not every instance of language use involves an intention to share information, however. Examples of this are "Stop tickling me!" (an intention to induce a certain behavior) or "Look, an eagle!" (an intention to attract attention, share an experience). To emphasise this point, Moore (2016c) reformulates the two intentions of the speaker as follows (conceded by Sperber & Wilson, 1995):

1. An intention to produce a particular response in the hearer/audience.
2. An intention that the hearer/audience recognizes intention 1. (Moore, 2016c)

Sperber and Wilson (1995) use the term *ostensive behavior* or simply *ostension* to describe communicative behavior that involves both these intentions; "behavior which makes manifest an intention to make something manifest" (1995, p. 49). To capture both the ostension on the side of the speaker and the inference on the side of the hearer in a unified model of pragmatics, they coined the term *ostensive-inferential communication*. This model describes the type of communication we find in humans, as opposed to communication systems that can be described by the code model. Other models of communication have also been proposed (e.g. Gärdenfors, 2003), but the contrast between the code model and the ostensive-inferential model of communication suffices to outline the questions that this article is concerned with.

At this point, there are two important things to note. Firstly, ostensive-inferential communication is something humans also do in non-linguistic communication. A tilt of the head or roll of the eyes are examples of ostensive behavior that can make the receiver look for an informative intention (such as "Look, Uncle Steve is getting drunk again")—and even completely novel, non-conventional gestures can be used to communicate ostensively, given that signaler and receiver share sufficient background. Second, the content of an informative intention can be recovered by a hearer even without recognizing the encompassing communicative intention. This is especially the case in non-linguistic and non-conventionalized ostensive behavior, such as moving someone's phone into their line of sight to make sure they don't forget it. The receiver of this signal may fulfil the signaller's goal even without realizing that the phone was moved there with an intention to signal something. However, the ability to recognize communicative intentions does make communication more efficient, because it points a hearer towards potentially relevant information. An act of ostension makes a receiver look for an informative intention—even if they do not directly see what the content of the informative intention is, recognizing that there is a communicative intention will motivate them to spend cognitive resources on inferring it (Csibra, 2010). This is what Sperber and Wilson (1995) refer to as the *principle of relevance*.

Although in theory this type of ostensive-inferential communication could be highly standardized and code-like (see e.g., Csibra, 2010), in practice, we see that humans can improvise ostensive signals on the fly and interpret utterances even if they are ambiguous and unexpected (Sperber & Wilson, 2002). This makes it highly likely that human communication involves some level of mental state attribution and thus theory of mind (ToM).[2] To answer the question of how this pragmatic competence *evolved,* however, we need a theoretical framework

for analyzing exactly what psychological processes are involved and what the precursors of these might be.

3. Psychological Mechanisms Underlying Pragmatic Competence

A good place to start when trying to identify the requirements for pragmatic competence and their precursors is Dennett's intentionality framework, which classifies the different levels of intentionality that can be ascribed to an organism (Dennett, 1983). A *zero-order* intentional system is, in fact, not an intentional system, because there are no mental states (such as beliefs and desires) behind the signal that the organism sends. The signal still counts as a signal however, because it is an adaptation that has evolved for the purpose of altering a receiver's behavior in a way that increases the sender's fitness (Maynard Smith & Harper, 1995). An example of this kind of signal is *aposematism* (warning coloration), which we find, for instance, in poisonous frogs that have evolved a salient skin color that warns predators of their toxicity: although this signal has a clear "message" for the predator ("Don't eat me"), there is no intentionality on the side of the signaler (Summers & Clough, 2001).

A *first-order* intentional system is an organism that, in the words of Dennett (1983), has beliefs and desires (etc.), but no beliefs and desires *about* beliefs and desires. For communication, this means that there is a mental representation underlying the signal, but no intention to modify another individual's mental state. Signals that are sent with such first-order intentionality are often referred to as *functionally referential signals*. This term was coined to accommodate the fact that, although signalers and receivers behave as if these signals refer to

specific objects or events in the same way that human words do, the mental processes underlying the production and reception of these signals may be very different from those involved in human language (Scarantino, 2013). The classic example of this type of signaling system are the alarm calls of vervet monkeys (although many species have similar systems of alarm calls). Vervet monkeys have different calls for different predators, and on hearing a call, group members will produce the corresponding evasive behavior (Seyfarth, Cheney, & Marler, 1980). However, current consensus is that these calls are most likely produced as a direct response to observing a predator, rather than with the intention to inform others (Zuberbühler, 2013); that is, they are more like a natural code than an instance of ostensive-inferential communication.

A *second-order* intentional system, then, is a system that also has beliefs and desires *about* the beliefs and desires of others. Dennett's orders of intentionality can go up even further (an example of *third-order* intentionality for instance is "Ella wants Steve to believe that she did not know about the surprise party"), and every order from second-order intentionality upwards involves the ability to entertain *metarepresentations*— to have representations *about* representations. This is something humans are remarkably good at—O'Grady, Kliesch, Smith, and Scott-Phillips (2015) showed that adults can keep track of mental state representations up to seven levels deep. How many levels of metarepresentation are minimally required to do ostensive-inferential communication is a question currently under debate, which is discussed in section 3.1, "Minimal Requirements for Ostensive-Inferential Communication."

All levels of intentionality exceeding first-order intentionality require an ability to represent the mental states of others (beliefs about beliefs) and thus a ToM. Levels of ToM can be counted in the same way as the orders of intentionality described above: first-order ToM is the ability to represent beliefs, second-order ToM is the ability to have beliefs about beliefs, etc. (e.g.,

Baron-Cohen, Jolliffe, Mortimore, & Robertson, 1997). A particularly well-studied kind of belief about belief is so-called *false belief* understanding (i.e., holding the belief that someone else has a belief that you know is not true). False belief understanding is special because it requires an understanding that other minds contain *representations* of the world that can be different from reality (Wellman, Cross, & Watson, 2001). It thus requires the individual to represent another's mental state in a way that is independent from their own representation of reality. As such, false belief understanding is often considered a hallmark of full-blown ToM capacity. In empirical studies of false belief understanding, a distinction is often made between explicit and implicit measures.

Explicit false belief understanding is measured in tasks where the participant has to give an explicit response based on their understanding of the false belief of another agent, for example, by pointing to or saying in which location a story character will look for a toy according to their false belief. This requires a capacity to overtly reason about others' mental states from a detached, third-person perspective (Helming, Strickland, & Jacob, 2014). Human children only start succeeding at these explicit tasks around the age of four (Wellman et al., 2001).[3] In contrast, *implicit* false belief understanding is measured using gaze direction or looking times, in tasks that don't require any explicit response or decision on the part of the participant. These tasks involve either measuring children's anticipatory looks to a location where they expect a story character will search based on the character's false belief, or the amount of time the child spends looking at the character when they search for their toy in the location that was *un*expected based on their false belief (with longer looking times indicating surprisal). This type of experiment has provided evidence that children are able to represent false belief-like states much earlier on, from as young as 7 months old (see Barrett et al., 2013; Southgate, Senju, & Csibra,

2007, for the anticipatory looking paradigm; and see Kovács, Téglás, & Endress, 2010; Onishi & Baillargeon, 2005; Surian, Caldi, & Sperber, 2007, for the violation-of-expectation paradigm).

Explanations of this discrepancy between when implicit and explicit false belief understanding become available can be divided into three kinds. First, there is the account that human infants are able to represent false beliefs from very early on (perhaps even from birth), but that the ability to produce the correct explicit response requires inhibition and selection mechanisms that take several years to mature. For instance, Leslie, Friedman, and German (2004) and Leslie (2005) argue that children have an innate mechanism for representing the mental states of others, but that they have learned as a default option that others' beliefs about the world are the same as their own (also known as a *reality bias; see also Birch & Bloom, 2004*). The development from an implicit to an explicit ToM ability then involves the development or maturation of a selection process that allows children to select among the different belief states they have represented; until this selection process is fully developed, children fail to suppress their reality bias, leading them to give the wrong answer in a false belief task.[4] This first account is thus compatible with the view that *explicit* false belief tasks do not accurately reflect the mindreading abilities of young children.

Second, there is the account that argues that it is the *representational* mechanism that has to mature, rather than the capacity to select between possible representations. For example, Rakoczy (2012) distinguishes between beliefs proper and *subdoxastic states*, which can be states like "has an inclination to think that" or "will be likely to behave as if she believes that.". A representation of a subdoxastic state such as "The character will have an inclination to think that the toy is in the yellow box" would produce the same results as a representation of the form "The character believes that the toy is in the yellow box," and the same is true for experiments using a

gaze direction or active helping paradigm. According to this account, subdoxastic states are different from proper beliefs because (a) they cannot be integrated with informational states from other areas of cognition, and (b) they are not accessible to conscious introspection, meaning that a child holding such representations would fail to produce the correct response in explicit (but not implicit) tests of false belief understanding. This second account is thus compatible with the view that *implicit* false belief tasks are not testing full-blown ToM ability.

Third, and finally, there is the two-systems account, which argues that implicit and explicit false belief tasks measure two separate systems that are both part of the full-blown human ToM capacity but that develop in different ways and at different ages. For instance, Apperly and Butterfill (2009) argue that later-developing, explicit false-belief understanding is a result of flexible cognitive processes that depend in their development on language and executive functions, whereas early, implicit false-belief understanding is a result of a set of less flexible, cognitively efficient processes that are available before language and executive functions develop. Given this hypothesis, Apperly and Butterill predict that early, implicit ToM is likely to be limited in rather arbitrary ways, both in terms of the type of content that can be represented (e.g., "that the toy is in the yellow box" vs. "that Ella doesn't know that Steve was not really ill") and the type of psychological roles that can be attributed (e.g., "x *believes* y" vs. "x *thinks* y" vs. "x *desires* y," etc.). This third account is compatible with the view that both implicit and explicit false belief tasks accurately measure some part of children's ToM, but that they tap into two different underlying systems.

When it comes to ostensive-inferential communication, there are two ToM abilities that have been argued as necessary: (a) the ability to entertain metarepresentations, and (b) the ability to represent beliefs (as opposed to subdoxastic states) (see Scott-Phillips, 2014; Sperber, 2000;

Sperber & Wilson, 2002; and Tomasello, 2008, for the metarepresentations claim; and see Breheny, 2006, for the beliefs claim). These arguments have subsequently been used to claim that this type of communication is unique to humans (Scott-Phillips, 2014; Sperber, 2000; Tomasello, 2008). However, in recent years, there have been moves to re-examine whether human communication necessarily involves such sophisticated mental operations, or whether the minimal cognitive requirements for doing pragmatics might be less demanding.

3.1 Minimal Requirements for Ostensive-Inferential Communication

For instance, Moore (2016a) argues that to understand informative communicative intentions, it is often sufficient to distinguish knowing from not knowing, and it is not necessary to have an understanding of false beliefs. To use an example of Moore (in turn adopted from Tomasello, 2008): if a sender makes a digging motion towards the ground to signal that there are likely to be tubers to dig for, this motion would be communicative in the original definition of Sperber and Wilson (1995) only when the sender has the intention to make the receiver *believe* that there are tubers in that patch of ground. However, Moore argues that for the sender to have the intention that the receiver should *attend to*, *see,* or *recognize* the presence of tubers would have the same effect and would make the signal no less communicative or intentional. Holding the intention that someone *attends to/sees/recognizes* the presence of tubers requires at most an ability to represent a registration or awareness relation between that individual and a piece of information, which is less cognitively demanding than representing a belief (i.e., a propositional attitude or representational relation that can be false) (Apperly & Butterfill, 2009; Martin & Santos, 2016). The same argument holds from the point of view of the receiver. Say if a fully ostensive sender has the intention to make the receiver *believe* (i.e., non-factual) that there are tubers in this

particular patch of ground, and the receiver understands this rather as the sender having the intention to make her *recognize* (i.e., factual) the presence of tubers, this will still produce the same behavioral response. This would also explain why infants seem to be able to recognize communicative intentions very early on (Csibra, 2010), without having to posit that they can already represent abstract mental states like beliefs[5].

Aside from an ability to represent beliefs, Sperber (2000) posits that ostensive-inferential communication also requires the ability to entertain *fourth-order* metarepresentations as the one depicted below (where *S* stands for sender and *R* stands for receiver):

fourth order:	*S* intends
third order:	That *R* believe
second order:	That *S* intends
first order:	That *R* believe
representation:	That there are tubers for which they could dig.

However, Moore (2016a) argues that ostensive-inferential communication consists of two *functionally distinct* components. The first component is the act of sending a signal (with the intention of invoking a certain behavioral response in the receiver), and the second component is the act of attracting attention towards the "signalhood" of that signal (with the intention of the receiver recognizing the first intention). Moore calls this first component the *sign production* and the second component the *act of address* (similar to the aforementioned act of ostension). Given this separation, we can break down the schema above into two separate second-order metarepresentations. For the act of address, the sender would (maximally) need to entertain a representation like the following:

second order:	*S* intends that

first order: *R* see that

representation: *S* is addressing to *R* an action *x*.

And for the act of sign production, a representation like the one below would suffice:

second order: *S* intends that

first order: *R* recognize that

representation: There are tubers for which they could dig.

This is already less cognitively demanding than the fourth-order metarepresentation analysis of Sperber (2000), but Moore (2016a) shows how in certain cases even lower-level metarepresentations would suffice. If for example the sign that the sender produces is a point to the ground, the second-order metarepresentations above could be reduced to first-order metarepresentations as follows:

Act of address:

first order: *S* intends that

representation: *R* attend and respond to her gesture.

Sign production:

first order: *S* intends that

representation: *R* looks at the ground by *S*'s feet.

Furthermore, from the perspective of the sender, it is not necessary to *explicitly* represent the first order of either of the above two metarepresentations. The sender only needs to *have* these intentions, she does not need to be aware of them.

To summarize, sending a signal ostensively and intentionally thus *minimally* requires only first-order metareprsentations in the case of *declarative* communication (i.e.

information-sharing) and no metarepresentations at all in the case of *imperative* communication (i.e. requests or demands, such as the pointing example above). From the receiver's perspective, there is always one extra level of metarepresentation required compared to what the sender needs to represent; second-order metarepresentations in the case of declarative communication, and first-order metarepresentations in the case of imperative communication.

Subdoxastic states and first-order metarepresentations could thus be potential precursors of the full-blown pragmatic competence we find in humans and could even turn out to be sufficient for doing some of our everyday linguistic communication. However, it seems likely that human language use *can* involve representations of proper belief states and fourth-order metarepresentations, at least sometimes. Before moving on to the question of how these representational skills have evolved, we will first review to which extent their precursors are present in other primates.

4. Pragmatic Competence in Great Apes

Comparative research is a good place to start when studying the evolution of a species-specific trait, because it offers valuable insights into the starting point from which the trait of interest evolved (Nunn, 2011). If precursors of the trait are present in related species, it is likely that those were already present in their last common ancestor with the species under investigation, and thus do not require a species-specific evolutionary account. In the case of human pragmatic competence therefore, the question we need to ask before theorizing about its evolution is what parts of this trait we share with other primates.

Here we will limit our discussion to the nonhuman great apes (i.e *Hominidae*—orangutans, gorillas, chimpanzees, and bonobos), because they are our closest living relatives and because most research on intentionality in nonhuman communication has focused on these species. We will first discuss the findings regarding ToM abilities in great apes, followed by the evidence that they employ these in their communication.

4.1 Mental State Representations

Most studies of great ape ToM have been conducted with captive chimpanzees. For instance, Kaminski, Call, and Tomasello (2008) explored ToM in a task in which two chimpanzees compete over food rewards. The chimpanzees were positioned opposite each other in separate enclosures, with a table with three cups placed in between them. In each trial, one of the chimpanzees (the subject) observed an experimenter placing food rewards in two of the three cups. The other chimpanzee (the competitor) either also witnessed the baiting of all cups, or of only one of them (in which case their view was occluded by an opaque panel during the baiting of one of the cups). Subsequently, both chimpanzees were allowed to choose one of the cups and receive its reward: either the subject got to choose first, or the competitor chose first (and the subject's sight was occluded while the competitor made their choice).

Kaminski et al. (2008) found that, when the competitor only saw the baiting of one of the cups, and the subject got to choose second, the subjects more often chose the unknown reward (the one not witnessed by the competitor) than the known reward. In contrast, when the subject was allowed to choose first, they were equally likely to go for the known and the unknown reward. Kaminski et al. concluded that chimpanzees can represent what others know based on what they have seen, and can predict their behavior accordingly.[6] The chimps behaved as if they knew that if the competitor only knows the location of one of the rewards, they are likely to pick

that one, which means that, when choosing second, the subject would be better to go for the reward that was unknown to the competitor. Kaminski et al. also conducted the same experiment with human children (mean age 6) and adults, and found a similar pattern of results.

In a second, *false* belief task, Kaminski et al. used the same set-up, but added a lift and a shift event, where, after the initial baiting of the cups, the reward was either lifted and replaced in the same cup (lift condition), or lifted and replaced in a different cup (shift condition). This lift or shift event was either witnessed by both participants, or by the subject only. In addition, Kaminski et al. now made the two rewards different in quality: one regarded as very desirable by both participants and one regarded as less desirable.

When running this experiment with six-year-old children, Kaminski et al. found that, in the condition where the subject got to choose second, they picked the high-quality reward more often than the low-quality reward in the unknown shift condition (where the shift had not been witnessed by the competitor) but not in the unknown lift condition. This shows that the children were able to distinguish between the condition where the competitor's belief about the high-quality reward was still accurate (unknown lift) and the condition where the competitor's belief had been rendered false (unknown shift). Chimpanzees on the other hand did not act differently in these two different conditions: in both cases they went for the cup containing the high-quality reward slightly more often than the cup with the low-quality reward.

Krachun, Carpenter, Call, and Tomasello (2009) elaborated on this study, using a similar competitive set-up but testing both chimpanzees and bonobos, and measuring looking times in addition to explicit choice responses to see if apes do show *implicit* signs of false belief understanding. In this study, there were only two cups and one reward, and the competitor was a human experimenter who either had a true or false belief about the location of the reward. The

human competitor got to choose first in each condition, but in the crucial trials, they intentionally did not manage to reach the cup in time before the table was moved over for the ape subject to make their choice. If the subjects were able to represent the competitor's false belief and predict her behavior accordingly, they could use the competitor's unsuccessful reach as an indicator of the reward's location (the reached-for cup in the true-belief case; the other cup in the false-belief case). As one would expect based on the results of Kaminski et al. (2008), the apes' explicit choice responses in these two conditions were not significantly different: in both cases they selected the reached-for cup (resulting in a reward in the true belief condition, and no reward in the false belief condition). Looking times, however, revealed a different pattern: subjects did *look longer* at the unchosen cup before making their choice in the false belief condition than in the true belief condition. This may indicate some awareness of the competitor's false belief, even if the subject was not able to use this for deciding which cup to choose. This could be either because these apes lack the necessary inhibition to suppress the tendency to go for the reached-for cup (an explanation in line with the failure-to-inhibit account of Leslie et al., 2004), or because their false belief representations are too subdoxastic to be integrated with the rest of their behavior-prediction procedures (following Rakoczy, 2012). When testing 4.5- to 5-year-old children on the same task, Krachun et al. (2009) found that they *did* respond as if they understood that the experimenter had a false belief: choosing the reached-for cup in the true belief condition and the other cup in the false belief condition.

More recently, Krupenye et al. (2016) looked specifically at great apes' *implicit* signs of false-belief understanding, using the eye-tracking method. In this study, the apes (chimpanzees, bonobos and orang-utans) watched videos of a human actor interacting with another actor in a King Kong costume. In one set of videos, the actor was looking for King Kong who was hiding

in one of two haystacks (experiment 1); in another set of videos, the actor was looking for a stone that King Kong had hidden in one of two boxes (experiment 2). In both experiments, King Kong rehid (himself or the stone) while the actor was in another room, in order to induce a false belief in the actor. The actor then returned and ambiguously approached both locations. During this ambiguous approach, the ape participants' first anticipatory look towards the two possible hiding locations was measured. Results showed that the apes made significantly more first looks towards the location where the actor falsely believed his target to be than to the 'true belief' location. This result, in accordance with the looking time results of Krachun et al. (2009), suggests that apes' abilities to understand beliefs may be similar to those of human infants.

Although caution should always be exercised in drawing conclusions from the relatively small number of studies that have been conducted on the ToM abilities of great apes, and absence of evidence cannot be taken as evidence of absence (especially not in primatology experiments, which are methodologically extremely challenging), we can tentatively conclude that great ape cognition includes the ability to represent mental states, but that these representations may fall short of proper beliefs that can be used to reason with and act upon.[7] As far as we are aware however, a study in the same vein as Kaminski et al. (2008) and Krachun et al. (2009) has not yet been run with human infants. Therefore, it is, as yet, unclear to what extent the difference in performance on these experiments between great apes and human children is due to a difference in biology and to what extent it is due to a difference in cultural input. Based on the current evidence, we can conclude that great apes have the beginnings of some of the cognitive capacities putatively involved in ostensive-inferential communication, but probably not at the same level of sophistication as seen in humans above the age of five. In addition, evidence has also been found that chimpanzees are able to entertain at least first-order metarepresentations

(Call and Carpenter 2001; Call 2010; Beran, Smith and Perdue 2013). These beginnings of belief understanding and metarepresentation may be just enough to fulfill the minimal requirements for ostensive-inferential communication as defined by Moore (2016a), described in the previous section.

4.2 Intentional and Ostensive Communication in Great Apes

A second, related question is to what extent great apes employ these ToM-like capacities in their communication. Most studies of primate pragmatics have focused on the question of whether great apes produce their signals (be it gestures or vocalizations) intentionally, that is, exhibiting an informative intention (with first- or second-order intentionality, according to the analysis of sign production in section 3.1). This is different from the question of whether great ape communication is ostensive, because ostension also requires a communicative intention, or in other words, overt intentionality. Liebal, Waller, Burrows, and Slocombe (2014, pp. 169-193) give an extensive overview of the different indicators of intentionality that have been adopted in studies of primate communication, and categorize some of these as strong and some as weak. The four weak criteria are (a) social use; (b) visual-orienting behavior or gaze alternation; (c) response-waiting; and (d) flexibility. The three strong criteria are (e) the production of a signal selectively for certain individuals in an audience (a subclass of social use); (f) the production of a signal only when the intended receiver is already attending to the signaler, or actively manipulating the attention of the receiver; and (g) persistence and elaboration of the signal when the communicative goal is not or only partially met. Active manipulation of the receiver's attention (part of criterion [f]) can also be viewed as an indicator of ostension, because it serves to draw attention to the fact that there is an informative intention; that is, it serves to signal the

signalhood (Scott-Phillips, 2015b). The same has been argued for eye contact (part of criterion [b]) (Gómez, 1994, 2007).

The most compelling evidence that great apes can have informative intentions when communicating comes from studies of chimpanzees' vocalizations. Elaborating on an experimental design by Crockford, Wittig, Mundry, and Zuberbühler (2012) using a model of a viper snake (a predator much feared by wild chimpanzees), Schel, Townsend, Machanda, Zuberbühler, and Slocombe (2013b) evoked alarm calls from chimps traveling in groups through the forest. They found that at least some types of alarm calls that the chimps produced in these episodes satisfy strong criteria for intentionality (criteria [e], social use; and [g], persistence), and one weaker criterion (visual-orienting behavior or gaze alternation, in that the alarm-calling chimp will alternate looking between the snake model and their audience). In a second study focusing on chimpanzees' food calls, Schel, Machanda, Townsend, Zuberbühler, and Slocombe (2013a) investigated whether these calls are directed at specific other individuals or not. The results of this study showed that feeding chimps were significantly more likely to produce rough grunts (a food call interpreted as a generic invitation to come and eat) for higher-ranking individuals and good friends than for others, and looked in the direction from which they expected the intended audience to appear while vocalizing.

These two studies provide the strongest evidence to date that non-human primates have something that looks like informative intentions in their natural communication. Informative intentions are, of course, only part of what it means to do ostensive-inferential communication, and the presence of informative intentions do not imply the presence of communicative intentions (see e.g., Bar-On, 2013). Instead, the best indicator for a communicative intention is ostensive behavior. The chimpanzees of Schel et al. (2013b) showed some of this in their

persistence behavior — an alarm-calling chimp would persist in alarm calling until their audience was safe — but to our knowledge no studies of primate communication have been conducted focusing specifically on ostensive behavior.

Moore (2016c) specifically reviews the possibility and occurrence of ostension in the gestural communication of great apes, and uses strong criterion [f] (deliberately solicit[ing] the attention of others before gesturing) as the indicator, citing two findings of such behavior. First, Povinelli et al. (2003) found that chimpanzees change the location of their gestures to make sure they are in the line of sight of a human experimenter. Second, Liebal et al. (2004) found that all four species of great apes moved into the line of sight of a human experimenter before gesturing to request food — chimpanzees and bonobos doing so even when they had to move away from the food in order to get in front of the experimenter. If moving oneself and one's gestures deliberately into the line of sight of an interlocutor is taken as an act of intentionally drawing the receiver's attention to the sign, these findings can be interpreted as acts of ostension.

Overall we can conclude that great apes do indeed use their limited understanding of mental states in their communication; producing signals with an informative intention (Schel et al., 2013b; Schel et al., 2013a) and showing some signs of ostensive behavior—at least in the case of captive apes communicating with human experimenters (Liebal, Call, Tomasello, Pika, Call, & Tomasello, 2004; Povinelli, Theall, Reaux, & Dunphy-Lelii, 2003).

5. The Biological Evolution of Human Pragmatic Skills

So far, we have seen that human pragmatic competence involves sophisticated ToM skills that allow humans to represent the beliefs of others in a way that is decoupled from their own representation of the world (e.g. Liu, Sabbagh, Gehring, & Wellman, 2004) and to entertain such

representations up to several levels of embedding (i.e., metarepresentations) (O'Grady et al., 2015). Our closest primate relatives (great apes) share some precursors of these skills, including the ability to represent what others know (Call & Tomasello, 2008; Kaminski et al., 2008; Krachun et al., 2009), perhaps some implicit awareness of beliefs (Krachun et al., 2009; Krupeney et al., 2016), and an ability to entertain at least first-order metarepresentations (Call and Carpenter 2001; Call 2010; Beran, Smith and Perdue 2013). Evidence has also been found that great apes put these abilities to use in their communication, both in captivity and in the wild (Liebal et al., 2004; Povinelli et al., 2003; Schel et al., 2013b; Schel et al., 2013a). Discussion is ongoing about whether or not this qualifies as ostensive communication proper (Moore, 2016c; Scott-Phillips, 2015b), but we will now turn to theories of how the *Homo* lineage got from this rather limited pragmatic competence to the pragmatic competence we find in humans today—specifically, the flexible usage of the ability to hold and recognize informative and communicative intentions, which allows for the use of highly ambiguous utterances and improvised ostensive signals. In the current section, we focus on explanations involving biological evolution, and in section 1.6, "The Cultural Evolution of Human Pragmatic Skills," we review explanations involving cultural evolution.

Biological evolution works with naturally occurring variation in traits that are transmitted genetically from generation to generation. The genes underpinning a particular trait are selected for if that trait increases the fitness (i.e., number of offspring) of an individual bearing that trait, relative to other competing traits. The best evidence that a trait has evolved by this route is, of course, to find the genes that code for the trait in question and to identify the signals of selection in their distribution, within and across populations. However, complex cognitive skills like those involved in ToM are probably reliant on many different genes interacting with each other and the

environment, making it hard to identify the genes involved (although see Xia, Wu, & Su, 2012, for a first attempt). As a result, other indicators are often used to try to work out if a given trait is genetically encoded and therefore potentially a target of natural selection, including: whether or not the trait in question comes online early on in infancy (indicating relatively little role for learning and therefore increasing the likelihood that the trait is largely determined genetically); whether it develops similarly in different individuals and different environments (again indicating a limited role for learning from experience); and whether there is a specialized neural substrate for the trait that can be selectively impaired (suggesting that the trait has relatively direct genetic underpinnings).

For ToM, the looking time studies of Onishi and Baillargeon (2005); Surian, Caldi, and Sperber (2007), and Kovács et al. (2010) suggest that infants are able to represent false belief-like states from as young as 7 months old, and a gaze-direction study of Barrett et al. (2013) suggests that implicit false-belief understanding in young children (1–4 years old) is similar across many different cultures. Together, these studies suggest that these capacities might be relatively experience-independent and therefore strongly constrained by genetics. In addition, neuroimaging studies of both typically developing adults and individuals with autism and other psychopathology suggest that humans have a brain network dedicated to ToM, which can be selectively impaired either from birth (as is the case in autism) or through brain injury later in life (see Brüne & Brüne-Cohrs, 2006, for a review). These neurological findings suggest that ToM has a relatively clear biological and genetic basis without which it cannot develop normally. However, cross-cultural studies of the developmental stages of mental state understanding (from 3 to 9 years old) show that cultural environment does have an influence, at least on the order in which different aspects of ToM are acquired (see Slaughter & Perez-Zapata,

2014, for a review). In addition, a twin study by Hughes, Jaffee, Happé, Taylor, Caspi, and Moffitt (2005) suggests that the majority of variance in ToM skills among individuals is explained by environmental rather than genetic factors. Thus, environment and learning contribute to ToM development as well.

Taken together, these observations suggest that at least some components of ToM are genetically transmitted and thus biologically evolved. Since these capacities seem uniquely well-developed in humans, this prompts the question of what selective pressures drove the elaboration of ToM and/or pragmatic capacities in our lineage—that is, what selective advantages would come from the ability to reason about the mental states of others?

Most accounts of how the biological underpinnings of pragmatic competence evolved in humans agree on the point that these evolved *before* language itself (i.e., the conventional code with vocabulary and grammar) existed (Csibra & Gergely, 2011; Scott-Phillips, 2014, 2015b; Sperber, 2000; Tomasello, 2008).[8] In this pragmatics-first view of language evolution, the ToM abilities that make up pragmatic competence initially evolved not for the purpose of language, but to serve some other function. Once this other pressure led to the improvement of ToM and/or metarepresentational abilities, these skills were then re-appropriated by language. Or, in the words of Scott-Phillips (2015b), language "is made possible by mechanisms of metapsychology and is made powerful by mechanisms of association" (Scott-Phillips, 2015b, p. 64) (where *mechanisms of association* refers to the ability to establish a conventional code where arbitrary vocalizations or gestures are associated with particular meanings, i.e., a vocabulary). This pragmatics-first account is reminiscent of the evolutionary process known as *exaptation*, where a particular trait gets co-opted for a use that is different from the one it was originally selected for (Gould & Vrba, 1982).

The question then becomes, why and how did the ToM abilities underlying pragmatic competence evolve, if it was not for language. Most theories that try to explain the remarkable social intelligence we find in primates, and humans especially, place its source in our increasingly complex social lives (e.g., Burkart, Hrdy, & Van Schaik, 2009; Byrne, 1996; Sterelny, 2012; Tomasello, Melis, Tennie, Wyman, & Herrmann, 2012; Whiten & Erdal, 2012). The advantages that full-blown ToM brings to such lives are an increased ability to predict and manipulate each other's behavior and an increased ability for cooperation. The hypothesis that human social cognition has evolved for the purpose of cooperation has been put forward by such as Sterelny (2012), Tomasello et al. (2012), and Whiten and Erdal (2012). The essential idea that these theories have in common is that there is something special about the hunter-gatherer lifestyle that hominins adopted during the Pleistocene, which made cooperation and honest information sharing beneficial enough to be selected for by biological evolution.

Because cooperating and sharing information are acts of trust that come at the risk of being exploited (e.g., Ale, Brown, & Sullivan, 2013), there are certain conditions that have to be met for cooperation to become adaptive (i.e., to constitute a selective advantage) (Sterelny, 2012). First, cooperation should come with a relatively high benefit and low cost. Second, individuals need to interact repeatedly to build up relations of reciprocal helping, allowing individuals to build up social alliances. Third, there should be a mechanism for detecting so-called free-riders (individuals who benefit without contributing). And finally, there should be a way of punishing these free-riders that is not too costly when compared to the benefits of cooperation. Sterelny (2012) and Tomasello et al. (2012) argue that these conditions were met when, due to a change in ecology, hominins in the Pleistocene started foraging collaboratively.

Collaborative foraging (such as big-game hunting) can only work if a group of individuals works together towards a joint goal and shares the spoils fairly.[9] Sterelny (2012), Tomasello et al. (2012), and Whiten and Erdal (2012) argue that this requires a ToM ability that is more sophisticated than what we find in great apes today, and that the selective advantage for (groups of) individuals who possessed such ability would have been strong enough for this trait to lead to more offspring. Aside from working together towards a joint goal, such improved ToM abilities would allow these early hominins to communicate more effectively; enabling them to work together on perfecting skills and tool use, and passing these on from generation to generation.[10]

This ability to pass on knowledge and skills from generation to generation by itself has also been argued to be the main selective pressure that has led to the sophisticated ToM ability and communication we find in humans. This idea is outlined in Csibra and Gergely's (2011) Natural Pedagogy hypothesis, which states that humans are born with a "well-organised package of biases, tendencies, and skills" (Csibra & Gergely, 2006, p. 8) that makes human infants particularly receptive to teaching. Specifically, this package includes the implicit ToM abilities that allow infants to recognize communicative intentions from very early on, through a special sensitivity to ostensive behavior (such as eye contact, infant-directed speech, and contingent reactivity) (Csibra, 2010). Csibra and Gergely (2011) argue that this natural pedagogy package is transmitted genetically, and that it evolved as a biological adaptation for teaching and cultural transmission. The argument here is that, as hominins developed skills and artefacts that became increasingly sophisticated and increasingly opaque in terms of their means-end relation, teaching became more and more important to enable reliable transmission of these skills and cultural practices. Such cultural transmission was important for evolving tool use and cooking practices,

which both had a clear selective advantage for humans (see respectively, Stout, 2011; Wrangham & Carmody, 2010).

To conclude, there may be certain ToM skills that have evolved specifically in humans because they formed biological adaptations to the hunter-gatherer lifestyle that our ancestors adopted during the Pleistocene. Two possible sources that gave rise to a selection pressure that resulted in abilities needed for ostensive-inferential communication are cooperation and cultural transmission, both of which benefit from an increased ability to represent intentions (both individual and shared) and to engage in ostensive communication. Interestingly, the second of these two adaptations—cultural transmission—in turn unlocks a much more rapid and flexible mechanism for adaptation: cultural evolution.

6. The Cultural Evolution of Human Pragmatic Skills

Many systems of human knowledge and behavior are culturally transmitted—passed on from generation to generation through social learning, rather than via genes. Cultural transmission leads to cultural evolution, where knowledge and skills accumulate over time, and adapt rapidly to the demands of both the environment and the minds through which they are transmitted (Henrich & McElreath, 2003). Humans are, by far, the most pervasively cultural species on the planet, and language (one of our many socially learned behaviors) is one of our most striking cultural feats (Smith & Kirby, 2008; Thompson, Kirby, & Smith, 2016). Could our unusually developed capacity for reasoning about mental states in others also be a product of cultural evolution?

Heyes (2012b) and Heyes and Frith (2014) review evidence from experimental, developmental, and neurocognitive studies showing that social learning plays a role in the

development of ToM, suggesting that ToM is (at least in part) a product of cultural evolution. First, as mentioned in the previous section, Hughes et al. (2005) found in a longitudinal twin-study that individual differences in mental state understanding are strongly correlated with verbal ability, and that this correlation is, for the most part, explained by environmental (rather than genetic) influences. In addition, Hughes et al. (2005) present indirect evidence that these environmental factors are composed largely of discourse with parents and siblings. Second, Heyes and Frith review several studies showing that children's ToM development is predicted by their parents' use of mental state terms and causal-explanatory statements about the mind (e.g., "She is smiling because she is happy") (Meins, Fernyhough, Wainwright, Das Gupta, Fradley, & Tuckey, 2002; Slaughter, Peterson, & Mackintosh, 2007; Taumoepeau & Ruffman, 2006; Taumoepeau & Ruffman, 2008). Third, the combined findings of Taumoepeau and Ruffman (2006) and Taumoepeau and Ruffman (2008) provide tentative evidence that parents (consciously or unconsciously) control their mental state discourse in such a way that they *tailor* it to the ToM abilities of their children. Taken together, these findings show a tight coupling between discourse about mental states and a child's ToM development.

In addition, Russell, Lyn, Schaeffer, and Hopkins (2011) compared great apes (chimpanzees and bonobos) who were reared in standard captivity environments (zoos and laboratories) to great apes reared in rich socio-communicative environments (ape language projects), to see how much influence socio-communicative training by humans could have on great apes' social cognition. The standard-reared apes in this study received only the necessary human interactions involved in feeding and other animal husbandry. The enculturated apes, on the other hand, had received extensive socio-communicative input from humans in the form of language training (training the comprehension of spoken language using specially designed lexigrams), although

not all apes included in the study had been equally successful at this task. The results of this study showed that, where the standard-reared apes performed worse on social cognition tasks (assessing communicative skills and understanding of attentional state and eye-gaze) than on physical cognition tasks, this difference was not present in the enculturated apes. Moreover, when compared to the performance of 2.5-year-old children on the same task, tested in a study by Herrmann, Call, Hernández-Lloreda, Hare, and Tomasello (2007), the enculturated ape group performed similarly to the children on the social cognition tasks, and even outperformed them on a task assessing understanding of the attentional state of an experimenter. Although the results of the standard-reared apes were not hugely different, they performed worse than the children on the task assessing the production of communicative signals and did not outperform the children in any of the other social cognition tasks. Similar results were found in a study by Lyn, Russell, and Hopkins (2010) looking at the ability of great apes to understand declarative signals (pointing and vocalizations). In this study, enculturated chimpanzees and bonobos were found to significantly outperform their standard-reared counterparts in their comprehension of ostensive points and vocalizations produced by human experimenters. The studies by Russell et al. (2011) and Lyn et al. (2010) thus show that environment can make a difference in the development of social cognition in great apes just as it does in humans.

This suggest a role for cumulative culture in the evolution of ToM. Although there might be a biological basis for ToM development that all humans share, the more sophisticated ToM abilities—such as higher-order metarepresentations and proper representational/propositional representations of mental states—may depend on cultural transmission. Heyes and Frith (2014) refer to these two parts of ToM as implicit and explicit ToM (echoing the conclusions of e.g., Kaminski et al., 2008; Krachun et al., 2009; Rakoczy, 2012). Implicit ToM skills in this

framework refer to the abilities responsible for the tracking of belief-like states found in infants by Onishi and Baillargeon (2005), Surian et al. (2007), and Kovács et al. (2010). These include gaze-following and joint attention, which develop early on in infancy and are shared with other great apes (and thus presumably part of our genetic endowment). Explicit ToM abilities, on the other hand, refer to that which allows humans to use their representations of the mental states of others *explicitly*, both in reasoning and behavior—this requires mental state representations that are independent from the individual's own representation of reality (i.e., so-called representational or propositional representations) (e.g., Apperly & Butterfill, 2009; Kampis, Somogyi, Itakura, & Király, 2013; Rakoczy, 2012). Based on the evidence summarized above, Heyes and Frith (2014) argue that these explicit ToM abilities develop through social learning rather than the maturation of innate cognitive modules.

As briefly mentioned, the power of cultural evolution is that it enables rapid accumulation of skills—where each generation can add some sophistication to the cognitive constructs that they get handed from the previous generation. In the case of explicit ToM abilities, this could take the form of increasingly elaborate, socially transmitted practices of discussing and reasoning about the mental states in others—also known as *folk psychology*. However, it is hard to imagine how such discussion and teaching about mental states would happen without language; especially considering the fact that all studies reviewed above as evidence for social learning of ToM place emphasis on the role of discourse with parents and siblings (Hughes et al., 2005; Meins et al., 2002; Slaughter et al., 2007; Taumoepeau & Ruffman, 2006, 2008). This leads to an interesting final hypothesis about the evolution of pragmatic competence: that ToM and language (in the sense of the conventional code with vocabulary and grammar) have *co-evolved*.

7. Have Language and Theory of Mind Co-Evolved?

The hypothesis that ToM and linguistic communication have co-evolved played at least some role in all theories of the evolution of human social cognition described in section 1.5 (Csibra & Gergely, 2011; Moore, 2016a; Sterelny, 2012; Tomasello et al., 2012; Whiten & Erdal, 2012;) and has been fleshed out more elaborately by Malle (2002). However, it is hard to find evidence for such scenarios of how cognitive skills evolved, since our ancestors in the *Homo* lineage have gone extinct and minds do not leave fossils. There are several types of indirect evidence that can be collected to test hypotheses like these (see e.g., Heyes, 2012a) however, one of which is evidence for co-development; if the development of one skill (e.g., explicit mindreading) is dependent on the development of another (e.g., language), the former could not have developed to the same extent when the latter had not yet evolved.

There is persuasive evidence consistent with the hypothesis that language and ToM co-develop. First, evidence that language-learning depends on ToM abilities is provided by Parish-Morris, Hennon, Hirsh-Pasek, Golinkoff, and Tager-Flusberg (2007). In a study comparing children with autism to typically developing children, they showed that, although 5-year-old autistic children have some ability to use social cues (pointing and eye gaze) to direct their attention in word learning, they perform at chance when learning new words requires inferring the speaker's intention, unlike language- and mental-age-matched typically developing children.

Second, the reverse phenomenon has also been observed, namely that the development of ToM depends in part on having access to language. Deaf children of hearing parents, who lack consistent linguistic input during the first years of their life, were shown to have delayed ToM development relative to deaf children of deaf parents, who receive sign language input from birth

(Schick, de Villiers, de Villiers, & Hoffmeister, 2007). Similarly, a study with typically developing children showed that simply training children on the use of mental state verbs with sentential complements accelerated their false belief understanding (Lohmann & Tomasello, 2003).

Third, in a study comparing different age groups of signers of the recently emerged Nicaraguan Sign Language, Pyers and Senghas (2009) showed that the bootstrap effect of language on ToM development continues on into adulthood. Pyers and Senghas found that the first cohort of signers (mean age 27), whose language had very limited mental state vocabulary, were worse at understanding false belief than the second cohort (mean age 17), who had more signs for mental states. Moreover, a follow-up study two years later revealed that the first cohort signers had improved in their false belief understanding, and that this either followed or co-occurred with, but never preceded, an expansion of mental state vocabulary.

Finally, a recent longitudinal study by Brooks and Meltzoff (2015) provides direct evidence that language and ToM co-develop. They showed that gaze following in 10.5-month-old infants predicted their production of mental state terms at 2.5-years, and that these mental state terms in turn predicted the extent of their false belief understanding at 4.5-years, even though gaze following did not directly predict false belief understanding. Thus, this shows evidence of an indirect relation between early sensitivity to social cues and later ToM ability, mediated by language.

Recent work by Woensdregt, Kirby, Cummins, and Smith (2016) has attempted to formalize this co-development hypothesis in a computational model in which Bayesian agents learn both a language and a way of inferring other agents' perspectives—replicating several of the co-development findings summarized above. How these co-developmental dynamics play out

over the course of (cultural and biological) evolution is an interesting question for future research that could be addressed with such a computational model, using the iterated learning framework (Kirby, Tamariz, Cornish, & Smith, 2015).

8. Biological, Cultural, and Co-Evolution of Pragmatic Competence

To conclude, pragmatics is a part of human language use that requires an evolutionary account of its own; separate from an account of how the linguistic code evolved. Pragmatic competence involves the ability to recognize and entertain informative and communicative intentions, which in turn requires an ability to represent mental states—often referred to as theory of mind (ToM). Although there are some ToM abilities that humans share with nonhuman primates—and that were already present before the linguistic code evolved—these abilities are limited in crucial ways when compared to the ToM abilities of adult humans. Specifically, nonhuman primates seem incapable of entertaining fully representational/propositional representations of mental states and are presumably also limited in their ability of entertaining higher-order metarepresentations.

One possibility is that these more sophisticated ToM abilities evolved in humans for the purpose of either cooperation or cultural transmission (or both), as a result of biological adaptation. Such biological evolution may have led to an increased sensitivity to acts of ostension and/or an increased motivation to engage in shared intentionality. However, another intriguing possibility is that (part of) these more sophisticated, explicit ToM abilities evolved

through cultural evolution—where cognitive skills are transmitted from generation to generation through social learning. This second possibility may have been unlocked by an initial biological adaptation that allowed for more reliable cultural transmission. Cultural evolution of ToM would have allowed for an accumulation of cultural practices for discussing and reasoning about the minds of others; which may have been key to the evolution of the sophisticated explicit ToM skills we find in humans today.

Such cultural accumulation of mental state reasoning may not have been possible without language however, which leads to the hypothesis that language (in the sense of the linguistic code) and pragmatic competence have co-evolved. This possibility deserves exploration in future research.

Further Reading

Ordered according to the structure of the chapter:

Books

Sperber, D., & Wilson, D. (1995). *Relevance: Communication and cognition.* Oxford, U.K.: Blackwell.

Scott-Phillips, T. C. (2015b). *Speaking our minds.* New York, NY: Palgrave Macmillan.

Wellman, H. M. (2014). *Making minds: How theory of mind develops.* Oxford , U.K.: Oxford University Press.

Apperly, I. (2010). *Mindreaders: The cognitive basis of "theory of mind."* New York, NY: Psychology Press.

Liebal, K., Waller, B. M., Burrows, A. M., & Slocombe, K. E. (2014). *Primate communication: A multimodal approach.* Cambridge, U.K.: Cambridge University Press.

Tomasello, M. (2008). *Origins of human communication.* Cambridge, MA: MIT Press.

Journal Articles

Scott-Phillips, T. C. (2015). Nonhuman primate communication, pragmatics, and the origins of language. *Current Anthropology, 56*(1), 56–80.

Moore, R. (2016c). Meaning and ostension in great ape gestural communication. Animal *Cognition, 19*(1), 223–231.

Moore, R. (2016a). Gricean communication and cognitive development. *The Philosophical Quarterly. Advance online publication.*

Tomasello, M., Melis, A. P., Tennie, C., Wyman, E., & Herrmann, E. (2012). Two key steps in the evolution of human cooperation: The interdependence hypothesis. *Current Anthropology, 53*(6), 673–692.

Heyes, C. M., & Frith, C. D. (2014). The cultural evolution of mind reading. *Science, 344(*6190), 1243091.

References

Ale, S. B., Brown, J. S., & Sullivan, A. T. (2013). Evolution of cooperation: Combining kin selection and reciprocal altruism into matrix games with social dilemmas. *PLoS ONE, 8*(5), 1–9.

Apperly, I. A., & Butterfill, S. A. (2009). Do humans have two systems to track beliefs and belief-like states? *Psychological Review, 116*(4), 953–970.

Bar-On, D. (2013). Origins of meaning: Must we "go Gricean"? *Mind & Language, 28*(3), 342–375.

Bar-On, D. (2016). Sociality, expression, and this thing called language. *Inquiry, 59*(1), 56–79.

Baron-Cohen, S., Jolliffe, T., Mortimore, C., & Robertson, M. (1997). Another advanced test of theory of mind: Evidence from very high functioning adults with autism or Asperger syndrome. *Journal of Child Psychology and Psychiatry, 38*(7), 813–822.

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a theory of mind? *Cognition, 21*, 37–46.

Barrett, H. C., Broesch, T., Scott, R. M., He, Z., Baillargeon, R., Wu, D., et al. (2013). Early false-belief understanding in traditional non-Western societies. *Proceedings of the Royal Society B, 280*(1755), 20122654.

Beran, M. J., Smith, J. D., & Perdue, B. M. (2013). Language-trained chimpanzees (*Pan troglodytes*) name what they have seen, but look first at what they have not seen. *Psychological Science*, *24*(5), 660–666.

Birch, S. A. J., & Bloom, P. (2004). Understanding children's and adults' limitations in mental state reasoning. *Trends in Cognitive Sciences*, *8*(6), 255–260.

Breheny, R. (2006). Communication and folk psychology. *Mind & Language, 21*(1), 74–107.

Brooks, R., & Meltzoff, A. N. (2015). Connecting the dots from infancy to childhood: A longitudinal study connecting gaze following, language, and explicit theory of mind. *Journal of Experimental Child Psychology, 130*, 67–78.

Brüne, M., & Brüne-Cohrs, U. (2006). Theory of mind, evolution, ontogeny, brain mechanisms, and psychopathology. *Neuroscience & Biobehavioral Reviews, 30*(4), 437–455.

Burkart, J. M., Hrdy, S. B., & Van Schaik, C. P. (2009). Cooperative breeding and human cognitive evolution. *Evolutionary Anthropology, 18*(5), 175–186.

Byrne, D. (1996). Machiavellian intelligence II. *Evolutionary Anthropology, 5*(5), 172–180.

Call, J. (2010). Do apes know that they could be wrong? *Animal Cognition*, *13*, 689–700.

Call, J., & Carpenter, M. (2000). Do apes and children know what they have seen? *Animal Cognition*, *4*, 207–220.

Call, J., & Tomasello, M. (2008). Does the chimpanzee have a theory of mind? 30 years later. *Trends in Cognitive Sciences, 12*(5), 187–192.

Carston, R. (2002). *Thoughts and utterances.* Malden, MA: Blackwell.

Crockford, C., Wittig, R. M., Mundry, R., & Zuberbühler, K. (2012). Wild chimpanzees inform ignorant group members of danger. *Current Biology, 22*, 142–146.

Csibra, G. (2010). Recognizing communicative intentions in infancy. *Mind & Language, 25*(2), 141–168.

Csibra, G., & Gergely, G. (2006). Social learning and social cognition: The case for pedagogy. In Y. Munakata & M. H. Johnson (Eds.), *Processes of change in brain and cognitive*

development: Attention and performance, XXI (pp. 249–274). New York, NY: Oxford University Press.

Csibra, G., & Gergely, G. (2011). Natural pedagogy as evolutionary adaptation. *Philosophical Transactions of the Royal Society B, 366*(1567), 1149–1157.

Dennett, D. C. (1983). Intentional systems in cognitive ethology: The Panglossian paradigm defended. *Behavioral and Brain Sciences, 6*, 343–390.

Gärdenfors, P. (2003). *How homo became sapiens: On the evolution of thinking.* New York, NY: Oxford University Press.

Gómez, J.-C. (1994). Mutual awareness in primate communication: A Gricean approach. In Parker, S., Boccia, M., and Mitchell, R., (Eds.), *Self-recognition and awareness in apes, monkeys and children* (pp. 61–80). Cambridge, U.K.: Cambridge University Press.

Gómez, J. C. (2007). Pointing behaviors in apes and human infants: A balanced interpretation. *Child Development, 78*(3), 729–734.

Gould, S. J., & Vrba, E. S. (1982). Exaptation: A missing term in the science of form. *Paleobiology, 8*(1), 4–15.

Grice, H. P. (1957). Meaning. *The Philosophical Review, 66*(3), 377–388.

Grice, H. P. (1975). Logic and conversation. In H. P. Grice (Ed.), *Studies in the way of words* (pp. 305–315). Harvard University Press.

Helming, K. A., Strickland, B., & Jacob, P. (2014). Making sense of early false-belief understanding. *Trends in Cognitive Sciences, 18*(4), 167–170.

Henrich, J., & McElreath, R. (2003). The evolution of cultural evolution. *Evolutionary Anthropology, 12(*3), 123–135.

Herrmann, E., Call, J., Hernández-Lloreda, M. V., Hare, B., & Tomasello, M. (2007). Humans have evolved specialized skills of social cognition: The cultural intelligence hypothesis. *Science, 317*(5843), 1360–1366.

Heyes, C. (2012a). New thinking: The evolution of human cognition. *Philosophical Transactions of the Royal Society B, 367*, 2091–2096.

Heyes, C. (2012b). What's social about social learning? *Journal of Comparative Psychology, 126*(2), 193–202.

Heyes, C. M., & Frith, C. D. (2014). The cultural evolution of mind reading. *Science, 344*(6190), 1243091.

Hughes, C., Jaffee, S. R., Happé, F., Taylor, A., Caspi, A., & Moffitt, T. E. (2005). Origins of individual differences in theory of mind: From nature to nurture? *Child Development, 76*(2), 356–370.

Kaminski, J., Call, J., & Tomasello, M. (2008). Chimpanzees know what others know, but not what they believe. *Cognition, 109*, 224–234.

Kampis, D., Somogyi, E., Itakura, S., & Király, I. (2013). Do infants bind mental states to agents? *Cognition, 129*, 232–240.

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

Kovács, Á. M., Téglás, E., & Endress, A. D. (2010). The social sense: Susceptibility to others' beliefs in human infants and adults. *Science, 330*(6012), 1830–1834.

Krachun, C., Carpenter, M., Call, J., & Tomasello, M. (2009). A competitive nonverbal false belief task for children and apes. *Developmental Science, 12*(4), 521–35.

Krupenye, C., Kano, F., Hirata, S., Call, J., & Tomasello, M. (2016). Great apes anticipate that other individuals will act according to false beliefs. *Science*, *354*(6308), 110–114.

Leavens, D. A., Russell, J. L., & Hopkins, W. D. (2010). Multimodal communication by captive chimpanzees (Pan troglodytes). *Animal Cognition, 13*(1), 33–40.

Leslie, A. M. (2005). Developmental parallels in understanding minds and bodies. Trends in *Cognitive Sciences, 9*(10), 459–462.

Leslie, A. M., Friedman, O., & German, T. P. (2004). Core mechanisms in "theory of mind." *Trends in Cognitive Sciences, 8*(12), 528–533.

Liebal, K., Call, J., Tomasello, M., & Pika, S. (2004). To move or not to move: How apes adjust to the attentional state of others. *Interaction Studies, 5*(2), 199–219.

Liebal, K., Waller, B. M., Burrows, A. M., & Slocombe, K. E. (2014). *Primate communication: A multimodal approach.* New York, NY: Cambridge University Press.

Liu, D., Sabbagh, M. A., Gehring, W. J., & Wellman, H. M. (2004). Decoupling beliefs from reality in the brain: An ERP study of theory of mind. *NeuroReport, 15*(6), 991–995.

Lohmann, H., & Tomasello, M. (2003). The role of language in the development of false belief understanding: A training study. *Child Development, 74*(4), 1130–1144.

Lyn, H., Russell, J. L., & Hopkins, W. D. (2010). The impact of environment on the comprehension of declarative communication in apes. *Psychological Science, 21*(3), 360–365.

Malle, B. F. (2002). The relation between language and theory of mind in development and evolution. In T. Givon & B. F. Malle (Eds.), *The evolution of language out of pre-language* (pp. 265–284). Amsterdam, Netherlands: John Benjamins.

Martin, A., & Santos, L. R. (2016). What cognitive representations support primate theory of mind? *Trends in Cognitive Sciences, 20*(5), 375–382.

Maynard Smith, J., & Harper, D. G. C. (1995). Animal signals: Models and terminology. *Journal of Theoretical Biology, 177*(3), 305–311.

Meins, E., Fernyhough, C., Wainwright, R., Das Gupta, M., Fradley, E., & Tuckey, M. (2002). Maternal mind-mindedness and attachment security as predictors of theory of mind understanding. *Child Development, 73*(6), 1715–1726.

Moore, R. (2014). Ontogenetic constraints on Grice's theory of communication. In D. Matthews (Ed.), *Pragmatic development in first language acquisition* (pp. 87–104). London, U.K.: John Benjamins.

Moore, R. (2016a). Gricean communication and cognitive development. *The Philosophical Quarterly.* Advance online publication.

Moore, R. (2016b). Gricean communication, joint action, and the evolution of cooperation. *Topoi.* Advance online publication.

Moore, R. (2016c). Meaning and ostension in great ape gestural communication. *Animal Cognition, 19(*1), 223–231.

Nunn, C. L. (2011). *The comparative approach in evolutionary anthropology and biology.* Chicago, IL: University of Chicago Press.

O'Grady, C., Kliesch, C., Smith, K., & Scott-Phillips, T. C. (2015). The ease and extent of recursive mindreading, across implicit and explicit tasks. *Evolution and Human Behavior 36*(4), 313–322.

Onishi, K. H.,& Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science, 308*(5719), 255–258.

Parish-Morris, J., Hennon, E. A., Hirsh-Pasek, K., Golinkoff, R. M., & Tager-Flusberg, H. (2007). Children with autism illuminate the role of social intention in word learning. *Child Development, 78*(4), 1265–1287.

Penn, D. C., & Povinelli, D. J. (2007). On the lack of evidence that non-human animals possess anything remotely resembling a "theory of mind." *Philosophical Transactions of the Royal Society B, 362*(1480), 731–744.

Povinelli, D. J., Theall, L. A., Reaux, J. E., & Dunphy-Lelii, S. (2003). Chimpanzees spontaneously alter the location of their gestures to match the attentional orientation of others. *Animal Behaviour, 66*, 71–79.

Pyers, J. E., & Senghas, A. (2009). Language promotes false-belief understanding: Evidence from learners of a new sign language. *Psychological Science, 20*(7), 805–812.

Rakoczy, H. (2012). Do infants have a theory of mind? *The British Journal of Developmental Psychology, 30*, 59–74.

Rubio-Fernández, P., & Geurts, B. (2012). How to pass the false-belief task before your fourth birthday. *Psychological Science, 24*(1), 27–33.

Russell, J. L., Lyn, H., Schaeffer, J. A., & Hopkins, W. D. (2011). The role of socio-communicative rearing environments in the development of social and physical cognition in apes. *Developmental Science, 14*(6), 1459–1470.

Scarantino, A. (2013). Rethinking functional reference. *Philosophy of Science, 80*(5), 1006–1018.

Schel, A. M., Machanda, Z., Townsend, S. W., Zuberbühler, K., & Slocombe, K. E. (2013a). Chimpanzee food calls are directed at specific individuals. *Animal Behaviour, 86*(5), 955–965.

Schel, A. M., Townsend, S. W., Machanda, Z., Zuberbühler, K., & Slocombe, K. E. (2013b). Chimpanzee alarm call production meets key criteria for intentionality. *PLoS ONE, 8*(10), 1–11.

Schick, B., de Villiers, P., de Villiers, J., & Hoffmeister, R. (2007). Language and theory of mind: A study of deaf children. *Child Development, 78*(2), 376–396.

Scott-Phillips, T. C. (2014). *Speaking our minds.* New York, NY: Palgrave Macmillan.

Scott-Phillips, T. C. (2015a). Meaning in animal and human communication. *Animal Cognition, 18*(3), 801–805.

Scott-Phillips, T. C. (2015b). Nonhuman primate communication, pragmatics, and the origins of language. *Current Anthropology, 56*(1), 56–80.

Seyfarth, R. M., Cheney, D. L., & Marler, P. (1980). Monkey responses to three different alarm calls: Evidence of predator classification and semantic communication. *Science, 210*(4471), 801–803.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal, 27*(3), 379–423.

Skyrms, B. (2004). *The stag hunt and the evolution of social structure.* Cambridge, U.K.: Cambridge University Press.

Slaughter, V., & Perez-Zapata, D. (2014). Cultural variations in the development of mind reading. *Child Development Perspectives,* 8(4), 237–241.

Slaughter, V., Peterson, C. C., & Mackintosh, E. (2007). Mind what mother says: Narrative input and theory of mind in typical children and those on the autism spectrum. *Child Development, 78*(3), 839–858.

Slocombe, K. E., & Zuberbühler, K. (2007). Chimpanzees modify recruitment screams as a function of audience composition. *Proceedings of the National Academy of Sciences, 104*(43), 17228–17233.

Smith, K., & Kirby, S. (2008). Cultural evolution, implications for understanding the human language faculty and its evolution. *Philosophical Transactions of the Royal Society B, 363*(1509), 3591–3603.

Southgate, V., Chevallier, C., & Csibra, G. (2010). Seventeen-month-olds appeal to false beliefs to interpret others' referential communication. *Developmental Science*, *13*(6), 907–12.

Southgate, V., Senju, A., & Csibra, G. (2007). Action anticipation through of false belief by attribution. *Psychological Science, 18*(7), 587–592.

Sperber, D. (2000). Metarepresentations in an evolutionary perspective. In Sperber, D. (Ed.), *Metarepresentations: A multidisciplinary perspective.* New York, NY: Oxford University Press.

Sperber, D., & Wilson, D. (1995). *Relevance: Communication and cognition* (2d ed). Oxford, U.K.: Basil Blackwell.

Sperber, D., & Wilson, D. (2002). Pragmatics, modularity and mind-reading. *Mind & Language, 17*, 3–23.

Sterelny, K. (2012). Language, gesture, skill: The co-evolutionary foundations of language. *Philosophical Transactions of the Royal Society B, 367*, 2141–2151.

Stout, D. (2011). Stone toolmaking and the evolution of human culture and cognition. *Philosophical Transactions of the Royal Society B, 366*(1567), 1050–1059.

Summers, K., & Clough, M. E. (2001). The evolution of coloration and toxicity in the poison frog family (Dendrobatidae). *Proceedings of the National Academy of Sciences, 98*(11), 6227–6232.

Surian, L., Caldi, S., & Sperber, D. (2007). Attribution of beliefs by 13-month-old infants. *Psychological Science, 18*(7), 580–586.

Taumoepeau, M., & Ruffman, T. (2006). Mother and infant talk about mental states relates to desire language and emotion understanding. *Child Development, 77*(2), 465–481.

Taumoepeau, M., & Ruffman, T. (2008). Stepping stones to others' minds: maternal talk relates to child mental language and emotion understanding. *Child Development, 79*(2), 284–302.

Thompson, B., Kirby, S., & Smith, K. (2016). Culture shapes the evolution of cognition. *Proceedings of the National Academy of Sciences, 113*(16), 4530–4535.

Tomasello, M. (2008). *Origins of human communication.* Cambridge, MA: MIT Press.

Tomasello, M., Melis, A. P., Tennie, C., Wyman, E., & Herrmann, E. (2012). Two key steps in the evolution of human cooperation: The interdependence hypothesis. *Current Anthropology, 53*(6), 673–692.

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development, 72*(3), 655–684.

Wharton, T. (2003). Natural pragmatics and natural codes. *Mind & Language, 18*(5), 447–477.

Whiten, A., & Erdal, D. (2012). The human socio-cognitive niche and its evolutionary origins. *Philosophical Transactions of the Royal Society B, 367*, 2119–2129.

Woensdregt, M. S., Kirby, S., Cummins, C., & Smith, K. (2016). Modelling the co-development of word learning and perspective-taking. In *Proceedings of the 38th Annual Meeting of the Cognitive Science Society*.

Wrangham, R., & Carmody, R. (2010). Human adaptation to the control of fire. *Evolutionary Anthropology, 19*(5), 187–199.

Xia, H., Wu, N., & Su, Y. (2012). Investigating the genetic basis of theory of mind (ToM): The role of catechol-o-methyltransferase (COMT) gene polymorphisms. *PLoS ONE, 7*(11).

Zuberbühler, K. (2013). Acquired mirroring and intentional communication in primates. *Language and Cognition, 5*(2–3), 133–143.

Notes:

[1] Although see *Consciousness and Cognition*, 2015, vol. 36 for a special issue on the extent to which certain mental states *can* be perceived directly.

[2] Note however that (Sperber & Wilson, 2002) propose that humans have evolved a "comprehension module" that is dedicated directly to inferring informative intentions once a communicative intention is recognized, which may reduce the amount of mindreading required.

[3] Although see Rubio-Fernández and Geurts (2012) for evidence that a different phrasing of the task allows children to pass it at three years old.

[4] See Helming, Strickland, and Jacob (2014) for a discussion of two other biases that may cause children to give the wrong response in explicit tasks (cooperative bias and referential bias).

[5] Although see Southgate et al. (2010) for evidence that 17-month-old infants *are* able to use belief representations to infer referential intentions.

[6] However see Penn and Povinelli (2007) for a discussion of alternative interpretations of such findings in terms of behavior rules.

[7] See Martin and Santos (2016) for another classification in terms of "awareness relations" versus "representational relations."

[8] Although see Bar-On (2016) for a different account, in which language and pragmatic ability evolved more in lockstep.

[9] What sets this type of foraging apart from the group hunting we see in lions and orcas, for example, is that collaborative foraging refers to a situation where (a) individuals *have to* collaborate in order to benefit; (b) the yield of a collaboration has to be greater than any solo foraging alternative; and (c) any alternative solo foraging has to be abandoned (risked) in order to collaborate. These three criteria are also what make up the "Stag Hunt" game in game theory (Skyrms, 2004).

[10] Although Moore (2016b) and others argue that ostensive-inferential communication does not require cooperation.

Chapter 3

Modelling the co-development of lexicon-learning and perspective-taking

In this chapter I will first provide an overview of empirical findings that inform us about the extent to which language-learning and perspective-taking (or mindreading more generally) are involved in each other's development. Section 3.1 reviews empirical evidence of how the ability to take a speaker's perspective (or 'read intentions') is important in the development of language, and Section 3.2 reviews evidence for the idea that language-learning in turn is important for the development of mindreading. This overview of empirical literature is followed in Section 3.3 by a review of computational models of word learning, focusing on those that make use of social cues and intention-reading to reduce referential uncertainty. The conclusions of this literature review motivate the development of a new model of word learning (presented in Section 3.4), in which the learner can use an ability to take the speaker's perspective in order to reduce referential uncertainty, but the speaker's perspective itself needs to be learned simultaneously with the lexicon. This model was previously published in the Proceedings of the 38th Annual Meeting of the Cognitive Science Society as Woensdregt et al. (2016), and the results presented in that proceedings paper are a precursor for the results presented in the current chapter. A pdf of Woensdregt et al. (2016) is reproduced in appendix F. Section 3.5 describes the developmental simulation results that were obtained with the model described in Section 3.4, showing that the simple incorporation of perspective-taking as a developing skill in a model of word learning

yields a co-developmental dynamic between these two skills. Finally, Section 3.6 discusses these simulation results in the light of the empirical literature reviewed in the first half of the chapter.

3.1 The role of mindreading in language development

Language is a culturally transmitted skill; it is acquired from others through social learning. Because languages consist for the most part of arbitrary form-meaning mappings and are learned in a noisy environment, language learners require some ability to attend to social cues and (eventually) to infer the referential and communicative intentions of others (e.g. Moore, 2016a). That these skills are important for acquiring language is evidenced by both typical and atypical language development. In this section I will first review empirical evidence of the role of intention-reading in typical language development (Section 3.1.1), followed (in Section 3.1.2) by a review of language development in children with autism spectrum disorder — of which one of the main characteristics is an impaired ability to infer others' mental states (Baron-Cohen et al., 1985; Baron-Cohen, 1995). In both literature reviews I will focus mostly on word learning, and within that on reference, as this provides a clear-cut and well-studied case of language development for which intention-reading is important. (I.e. the challenge faced by word learners is to pick out the speaker's intended referent among a number of potential referents.) This does not mean however that mindreading does not play a role in other areas of language development as well, including more abstract semantics, phonology, morphology, syntax, and of course pragmatics, as I will briefly discuss in Section 3.1.2.

3.1.1 Inferring speakers' communicative intentions for word learning in typically developing children

Several controlled experiments have shown that typically developing children are able to use social cues (such as the speaker's eye gaze) and make inferences about a speaker's referential intent in order to determine the referent of a novel word (see Baldwin and Moses, 2001; Tomasello, 2000, for reviews). To test infants' ability to use a speaker's eye gaze in word learning, Baldwin (1991, 1993b) developed a word learning task with two different conditions. In both conditions the experimenter was holding one of two novel toys while the child was holding the other, and the experimenter waited for a

moment in which the child focused her gaze on her 'own' object. In the *follow-in labelling* condition, the experimenter then also looked at the child's toy and labelled it with a novel word. In the *discrepant labelling* condition, the experimenter instead looked at their own toy while uttering the new label, so that experimenter and child were each looking at a different toy at the moment of labelling. The combined results of Baldwin (1991) and Baldwin (1993b) showed that at least from the age of 18 months onwards, infants are able to establish the 'correct' word-object mapping (novel label goes with the object that the speaker is looking at at the time of utterance) not only in the *follow-in* condition, but also in the *discrepant* labelling condition. These results suggest that at least from the age of 18-months-old, infants are able to use the speaker's eye gaze to guide them in learning word-object mappings.

Although these experiments demonstrate that infants are able to use eye gaze as a cue to reference, they leave open the question whether they do this on the basis of an understanding of referential intentions, or whether they use eye gaze as a more low-level heuristic. According to the former, 'rich', interpretation, infants use eye gaze to guide them in learning word-object mappings because they have an implicit understanding that both labels and nonverbal cues reflect what is on the speaker's mind. In contrast, the latter, 'frugal', interpretation can come in several possible forms. One of these (put forward by Baldwin, 1991) is that infants use eye gaze as a cue to word reference because they have learned that following someone's eye gaze can lead to interesting visual experiences, and (separately) that eye gaze and labels are often directed towards the same object. Importantly, for this account to work, infants do not need to have an understanding of referential intentions. That is, they do not need to understand *why* there is a relationship between the speaker's eye gaze and their utterances. Another low-level account (put forward by Baldwin, 1993a) could be that the speaker's eye gaze increases the salience of an object for infants, and that they simply link the novel label to the object which is currently most salient to them.

Baldwin (1993a) devised a new set of experiments to tease apart the rich interpretation from the second of these frugal accounts (the one based on salience). In a first experiment, Baldwin (1993a) addressed infants' ability to ignore temporal contiguity. Here, 19-20 month-old children were assigned either to a *coincide* or a *conflict* condition. In both conditions, the infants were first familiarised with two novel toys, which the experimenter then placed in two separate opaque containers, without the infant being able to see which toy went in which container. The experimenter then opened one

of the containers and uttered a novel label while looking intently into it, but without the toy being visible to the infant. In the *conflict* condition, this was followed by the experimenter first giving the toy from the *other* container to the infant to explore, after which she waited for at least 10 seconds before opening the container with the labelled toy and handing that toy to the child. In the *coincide* condition, the experimenter instead gave the labelled toy to the infant first, and waited for at least 10 seconds before handing them the other toy. Subsequent comprehension tests showed that infants mapped the novel label to the first toy that was given to them significantly more often in the *coincide* condition than in the *conflict* condition. Furthermore, infants in the *conflict* condition made fewer mapping errors (i.e. mapping the label to the first toy that was given to them) than would be expected by chance. Taken together, these results indicate that infants assign higher importance to eye gaze as a cue to reference than they do to temporal contiguity. However, the question remains whether infants do this because they understand eye gaze to be an index of the speaker's referential intent, or whether the speaker's eye gaze simply increases the salience of an object and infants map a novel word to the most salient object.

Baldwin (1993a) ran a second experiment to tease apart these two possible interpretations. This experiment was identical to the first, except that now the experimenter looked at the infant while uttering the novel label, instead of into the container she was holding. She did however lift one of the containers and adjust its lid during labelling, in order to enhance its salience. The results of a comprehension test showed that infants in this experiment did not establish any consistent word-object mappings. That is, their responses did not differ from what would be expected by chance. To check whether the experimenter holding and adjusting the lid of the target container really worked to enhance its salience, Baldwin (1993a) also measured the amount of time the infants spent looking at the target container during labelling in the second compared to the first experiment, and found no difference.

In sum, the results of Baldwin's (1993a) study indicate that infants' word learning behaviour as described above is based on an understanding of reference that goes beyond mere temporal contiguity or enhanced salience. Similar results were found by Baldwin et al. (1996), who compared a condition in which a novel label was uttered while only one toy was present and the speaker was visibly paying attention to it, with a condition that was identical except that the speaker was hidden from the infant's view. Baldwin et al. (1996) found that 18-20 month-old infants only mapped the novel label to the

toy when the speaker was visible, providing further evidence for the theory that infants treat social cues as uniquely relevant in the process of word learning.

For slightly older children (24-month-olds), Tomasello and Barton (1994) found even stronger evidence that they use an understanding of referential intent to guide their word learning. In Tomasello and Barton's Study 4, an experimenter stated her intention to find a toy which she described repeatedly using a novel label, and then started searching in five opaque containers which each contained one novel toy. In the *with search* condition, the experimenter first extracted the toys from the first two containers and rejected them (by frowning and putting them back), before finding the target toy in the third container, expressing glee and handing it to the child. In the *without search* condition, the experimenter simply immediately found the target toy in the first container she opened. Tomasello and Barton then tested the children on both elicited production trials (where the experimenter held up the target toy and asked "what's this?") and elicited comprehension trials (where the experimenter placed the five toys in front of the child and asked "Can you give me the *toma*?").

If the children had been using temporal contiguity rather than the experimenter's social cues, they should choose the object that was taken out of the first container to be the referent of the novel label in both conditions. Only one of the 15 participants in the *with search* condition did so, and out of the remaining 14 participants, eight correctly mapped the novel label to the target object. Moreover, Tomasello and Barton found no difference in children's production and comprehension behaviour between the *with search* and *without search* conditions. Taken together, these results indicate that the children used their understanding of the speaker's referential intent (based on her expressing her intention to look for the *toma*) to direct them in learning the referent of the novel label. In their Study 3, Tomasello and Barton showed similar results for children learning a verb for an intentional action, which required them to ignore an 'accidental' action.

Akhtar and Tomasello (1996) devised variations on these experiments in which two-year-old children never got to see the target object or target action of the novel label after the label was introduced (until test). The children were familiarised with the target object or action through several rounds of scripted play with the experimenter. But when the experimenter first uttered the novel label, together with her intention to either find the target object or watch a toy character perform the target action, she either found that the toy barn which contained the target object was 'locked', or

that the toy character who had previously performed the target action 'went missing'. Despite not seeing the target object or action at all after the novel word was uttered, children in both experiments still made the correct inference about the word's referent.

Tomasello et al. (1996) replicated both Tomasello and Barton's (1994) Study 4 (in which the experimenter searched and rejected several objects before finding the target object), and Akhtar and Tomasello's (1996) Study 1 (in which children did not get to see the target object after labelling because the toy barn was 'locked') with slightly younger children of 18 months old, and found similar results, showing that also 1.5-year-olds can learn novel nouns under these circumstances.

Eye-tracking can be used to test what word-object mappings infants have established at even younger ages, using the preferential looking paradigm (e.g. measuring the proportion of looks to a target item relative to a distractor item). This approach doesn't require the infants to make an explicit response (such as pointing to or selecting the target object out of an array of distractors, performing the target action, or verbally producing the learned word, as were used in the studies described above). In addition, eye-tracking can be used to shed light on how infants make use of a speaker's eye gaze during a labelling event. Yurovsky and Frank (2017) used this approach to investigate children's gaze following in simple word learning experiments across six different age groups from 1 through to 3.5 years old (i.e. each group spanned a six-month age range). In this study, children were first trained on novel word-object mappings by watching videos of a speaker seated at a table in between two novel toys, who then turned to one of the toys and labelled it with a novel word. After this training phase, children were tested on whether they had learned the correct mapping using the preferential looking paradigm.

In their Experiment 1, Yurovsky and Frank (2017) made sure the two novel toys did not differ in their visual salience, using measures collected in a separate looking time experiment. The results of this first experiment showed that children of all age groups followed the speaker's gaze to her intended referent, but that this ability improved over developmental time. Specifically, children became better with age at disengaging from the speaker's face (which all age groups spent most of their time looking at during training) in order to look at the target object. Combining this with the results of the test trials revealed that better gaze following to the target object went together with improvements in children's ability to pick out the intended referent during test.

In their Experiment 2, which included only the three youngest age groups (1-1.5,

1.5-2, and 2-2.5 years old), Yurovsky and Frank (2017) manipulated the visual salience of the two toys, thereby creating a *Salient* condition in which the target of labelling was also the salient toy, and a *NonSalient* condition in which target and competitor were swapped. When perceptual salience was in conflict with the speaker's eye gaze, children's looking behaviour at test showed no signs of them having learned any word-object mapping (i.e. neither between the word and the target, nor between the word and the competitor). Thus, although children from as young as 1 year old are able to use the speaker's eye gaze as a cue to reference, this can be disrupted if this cue is in conflict with perceptual salience (at least up to the age of 2.5 years).

Also using eye-tracking, Nappa et al. (2009) showed that slightly older children (divided in groups of 3, 4, and 5 years old) were able to use a speaker's eye gaze in order to learn the meaning of a novel verb in an ambiguous context. In their Experiment 1, children were shown images of ambiguous action events such as chasing/fleeing, which could be described as both "Character A is chasing Character B", and "Character B is fleeing Character A". These images were presented together with a video of a speaker looking at the scene and uttering an ambiguous sentence of the form "He's *blicking* him". Importantly however, the speaker first shifted his gaze to one of the two characters, and held it there while uttering the sentence. Because the perspective of these types of actions (i.e. whether the verb describes the action from the perspective of the instigator or the recipient) is not intrinsic to the event, but instead exists in the mind of the speaker, a child learning the meaning of these types of verbs has to somehow infer the speaker's communicative intention. One way in which children might solve this problem is by following the speaker's eye gaze. Previous studies and a separate experiment run by Nappa et al. (2009) showed that both children and adults have a bias in favour of interpreting these ambiguous verbs from the perspective of the instigator. Therefore, Nappa et al. were specifically interested in whether children could overcome this conceptual bias on the basis of an eye-gaze cue from the speaker.

After watching these training videos, children in the Nappa et al. (2009) study were asked to describe what the novel verb meant. Results showed firstly that, as predicted, children had a general preference for interpreting the verb from the instigator's perspective. However, when the speaker looked at Character B (the recipient) before uttering the sentence, children more often interpreted the verb from the recipient's perspective (e.g. 'fleeing') than from the instigator's perspective. Both effects held across all three age groups. Eye-tracking measures collected while children were watching the videos

showed that they followed the speaker's gaze towards one or the other character, thereby aligning their perspective on the event with that of the speaker. Taken together, these results show that children can use a speaker's gaze not just to make inferences about an intended referent, but also to make inferences about what a sentence is 'about' (i.e. how the speaker is framing the event).[1]

In sum, the experiments reviewed in this section show that infants and young children pay attention to social cues about the speaker's communicative intentions when learning words. The findings also suggest that children assign these social cues a privileged role in word learning compared to contextual cues such as temporal contiguity. Moreover, infants may refrain from establishing any word-object mappings if they cannot monitor the speaker for social cues (such as when the speaker is hidden from view in Baldwin et al., 1996). For typically developing children, learning about words is thus an intrinsically social matter, and is treated differently from learning about other regularities in the environment. However, findings from typically developing children alone cannot answer to what extent the ability to infer communicative intentions is *necessary* for language development. This question is addressed by the language development of children with autism spectrum disorder, which is discussed in the next section.

3.1.2 Language development in children with autism

That the ability to infer communicative intentions is important for acquiring language is also evidenced by the language development of children with autism spectrum disorder (ASD). ASD is a developmental disorder characterised by impaired social interaction and communication, as well as restricted interests and repetitive behaviours (DSM-V, 2013). One of the most well-known components of ASD is an impaired ability to understand social cues and infer others' mental states (Baron-Cohen et al., 1985; Baron-Cohen, 1995). Language development in individuals with ASD varies widely in terms of its trajectory and outcome (Anderson et al., 2007; Pickles et al., 2014), but difficulties with pragmatics and discourse are virtually universal (Eigsti et al., 2011; Tager-Flusberg et al., 2005). A significant delay in the development of language is also very common in children with ASD, who produce their first words around 38 months, compared to

[1]Nappa et al. (2009) contrasted this experiment with a second experiment in which the speaker's utterance also provided syntactic information about the verb's meaning (e.g. "The rabbit's *blicking* the elephant"). Results of the trials in which this linguistic cue was in conflict with the eye-gaze cue showed that children of all age groups assigned more weight to the linguistic cue. A non-significant trend in the data indicated that this preference for the linguistic cue became stronger over developmental time.

8-14 months in typically developing (TD) children (Eigsti et al., 2011; Howlin, 2003). Delayed or atypically developing speech is in fact often the first symptom leading to referral of children later diagnosed with ASD (De Giacomo and Fombonne, 1998), and about 25-30% of diagnosed individuals never go on to acquire functional speech (Anderson et al., 2007; Pickles et al., 2014; Wodka et al., 2013) Fluency and flexibility of expressive language are also two of the dimensions contributing to the distinction between 'low-functioning' and 'high-functioning' autism (Tager-Flusberg et al., 2005).

Language atypicalities in individuals with autism

As mentioned above, the most universally affected domains of language in individuals with ASD are pragmatics and discourse (Eigsti et al., 2011; Tager-Flusberg et al., 2005); these are the aspects of language that are most reliant on social considerations, such as the knowledge, interests, motivations and social status of the interlocutor. Impairments in the domain of pragmatics include difficulty with the production and comprehension of referential expressions that rely on taking into account the listener's knowledge and perspective, such as pronouns and deictic terms (Hobson et al., 2010a,b; Novogrodsky, 2013), and difficulty using the context to disambiguate utterances (Loukusa et al., 2007). Impairments in the domain of discourse include impaired narrative ability, as measured in the form of macrostructure (coherence and cohesive adequacy), microstructure (length of utterances and lexical diversity) and the use of internal state language to explain the motivations of characters in the narrative (see Stirling et al., 2014 for a review; and Baixauli et al., 2016 for a meta-analysis). Pragmatics and discourse are impaired in individuals with low- and high-functioning autism alike, and residual weaknesses in these domains have been shown to remain even in individuals who were diagnosed on the autism spectrum during their preschool years but who, thanks to early intervention, later no longer meet the criteria for diagnosis (Kelley et al., 2006).

In addition to impairments in pragmatics and discourse, prosody is also often found to be affected in individuals with ASD (including pragmatic prosody; Tager-Flusberg et al., 2005). Furthermore, evidence has been found of delays and deficits in phonology, morphology, syntax, and semantics (see Eigsti et al., 2011, for a review). One area of semantics that has been consistently found to be impaired is the comprehension and expression of mental state verbs like *think*, *guess*, and *know* (Kazak et al., 1997; Kelley et al., 2006; Tager-Flusberg, 1992; Ziatas et al., 1998), supporting the theory

that the symptoms of ASD (including language difficulties) are caused by an impaired mindreading ability. However, it is important to note at this point that this *Theory of Mind Deficit* account of ASD (Baron-Cohen, 1988; Surian et al., 1996) is not the only theory of autism which has been put forward to explain impairments in pragmatics and discourse. Two other influential theories of ASD, the *Weak Central Coherence Hypothesis* and the *Executive Dysfunction* theory provide alternative accounts for these impairments (see Martin and McDonald, 2003, for a review).

The Weak Central Coherence Hypothesis is an account not just of the cognitive weaknesses but also the strengths found in individuals with ASD, such as exceptional factual knowledge and attention to detail. This hypothesis thus characterises autism as a cognitive *style*, rather than a deficit (Happé and Frith, 2006). Happé and Frith use the term *central coherence* to describe an information-processing style which focuses on globally coherent patterns of information and meaning in context, at the expense of attention to detail. Weak central coherence, in contrast, is a detail-focused processing bias resulting from superior processing of local information, possibly at the cost of finding contextual meaning. According to this theory, the cognitive style of individuals with ASD is characterised by such weak central coherence (Happé and Frith, 2006; see also Frith and Happé, 1994 and Happé, 1999 for earlier versions of this account). Out of all domains of language, pragmatics and discourse rely most heavily on taking into account the broader context of conversation[2]. Therefore, the Weak Central Coherence hypothesis naturally predicts that individuals with ASD would experience specific difficulties in these domains (Martin and McDonald, 2003). According to this account, the deficits in pragmatics and discourse are thus not caused by a specific social-inference deficit, but by a broader difficulty with using context to derive meaning.

The Executive Dysfunction theory of autism has been put forward mainly to account for the restricted interests and repetitive behaviours found in individuals with ASD (Hill, 2004; Ozonoff et al., 2004; Pennington and Ozonoff, 1996). Executive functions is an umbrella term for the cognitive processes that are responsible for purposeful, goal-directed activity, such as planning, working memory, impulse control, inhibition and mental flexibility. The Executive Dysfunction theory of autism states that these processes are impaired in people with ASD (see Hill, 2004, for a review). Individuals with impaired executive functions may show deficits in pragmatics and discourse

[2]The pragmatic function of prosody can also be added to this list, and as mentioned above this is indeed often found to be impaired in individuals with ASD (Tager-Flusberg et al., 2005).

because they have difficulty integrating and responding to incoming information from multiple sources, because the 'rules' of conversation change with each context, or because they have difficulty inhibiting inappropriate responses (Eigsti et al., 2011; Martin and McDonald, 2003).

These three different theories of autism are not mutually exclusive. In fact, Happé and Frith (2006) argue that in the light of evidence of independence of the specific symptoms that the three accounts explain, the most plausible hypothesis is to view ASD as a result of anomalies in several different core cognitive processes, including mindreading, central coherence, and executive functions. Happé et al. (2006) come to the same conclusion on the basis of reviewing evidence from genetic, developmental and (neuro)cognitive studies. If it is the case that several of these three theories of ASD have independent explanatory power, a compounding of the effects of each might explain why pragmatics and discourse are almost universally affected across individuals with ASD, despite wide-ranging individual differences in other aspects of language. Namely, all three accounts predict difficulties in pragmatics and discourse, whereas their predictions diverge for other domains of language and cognition.

Word learning in children with autism

Turning now to the *delays* in language development found in children with ASD, these may be explained in part by an atypical strategy for word learning. Building on the work by Baldwin described in Section 3.1.1 above, Baron-Cohen et al. (1997) investigated whether children with ASD have more difficulty using social cues for word learning than TD children do. For this purpose, Baron-Cohen et al. replicated Baldwin's (1993b) experiment (contrasting *follow-in labelling* and *discrepant labelling*) with a group of autistic children (comparing their results to both TD children and children with learning disabilities matched to the ASD group on verbal mental age). Baron-Cohen et al. found that whereas the ASD group performed above chance on the *follow-in* condition (no different from the children with learning disabilities), they performed at chance on the *discrepant labelling* condition (significantly worse than the learning disability group). The children with ASD tended to map the novel word to the toy that their own gaze was fixated on during labelling, regardless of the target of the speaker's gaze. These findings suggest that children with ASD have difficulty learning words when it requires paying attention to social cues such as eye gaze.

However, Baldwin and Moses (2001) pointed out several caveats in Baron-Cohen et al.'s (1997) study and conclusions. Firstly, the autistic children's difficulties in the *discrepant labelling* condition could stem from more general information-processing difficulties, rather than a specific social-inference deficit (see discussion of the Weak Central Coherence Hypothesis and the Executive Dysfunction theory above). Secondly, we don't know whether what caused the difficulties for the ASD group was a failure to monitor the speaker's eye gaze at all, or a failure to use eye gaze information in the appropriate way to infer the correct mapping. Both of these criticisms are addressed to some extent by Preissler and Carey (2005), who not only replicated the Baron-Cohen et al. (1997) study, finding similar results, but also ran a follow-up experiment in which instead of social cues, children had to use the mutual exclusivity principle (Markman and Wachtel, 1988) to correctly infer the referent of a novel word. Preissler and Carey showed that the same group of children with ASD who failed their *discrepant labelling* condition, succeeded at this mutual exclusivity experiment (not performing any differently from a control group of TD toddlers).

These results suggest that at least to the extent necessary for using the mutual exclusivity principle, children with ASD are able to integrate information appropriately. Preissler and Carey (2005) also recorded how often each of the children looked at the experimenter's face during labelling, and reported that 19 out of 20 TD children did so, versus 3 out of 18 children in the ASD group. This finding indicates that for most children in the ASD group, things already started going awry at the stage of collecting the relevant eye gaze information. Preissler and Carey's mutual exclusivity condition was later criticised by de Marchena et al. (2011), because their experimental design confounded exclusivity with novelty. However, using a new design that eliminated this confound, de Marchena et al. found the same result in terms of children with ASD being able to use the mutual exclusivity principle in word learning.

In a study looking at 3-year-olds with an increased risk of autism (based on the fact that they have an older sibling with ASD), Gliga et al. (2012) showed that although following the speaker's gaze is necessary for word learning, it is not sufficient. Children at risk of autism often display subclinical ASD-like atypicalities (the so-called 'broader autism phenotype') even if they do not go on to be diagnosed with ASD. Gliga et al. found that those children in their sample who were categorised as having 'poor skills' according to a standardised test of social and communication atypicalities, performed at chance on a word learning task which required them to map a novel label to an

object that the speaker was looking at, instead of a more salient distractor object. (The 'poor-skills' group performed significantly worse than both their 'typical-skills' at-risk counterparts, and a control group of TD children with no family history of autism; both of which performed above chance.) The children in the 'poor-skills' group *did* however follow the speaker's eye gaze during labelling; not significantly less so than the two control groups. Gliga et al. interpret these findings as illustrating that there is a difference between 'gaze following' and 'gaze-reading', and that both are necessary for successful word learning.

In addition to the two caveats pointed out by Baldwin and Moses (2001), Luyster and Lord (2009) criticise the fact that Baron-Cohen et al. (1997) and Preissler and Carey (2005) focus on those autistic children in their studies who used the 'listener's direction of gaze' strategy (leading to the incorrect mapping in the *discrepant labelling* condition), while each of these studies also included some autistic children who used the 'speaker's direction of gaze' strategy on several occasions. Luyster and Lord (2009) therefore replicated the *follow-in* versus *discrepant* labelling experiment with a group of autistic children who were selected using an entry task in which they were asked to select an object of which they knew the label well. In addition to this inclusion criterion, Luyster and Lord also added extra contextual supports to their experiment, consisting of (i) repeating the novel label nine times instead of two (as was the case in Baron-Cohen et al. (1997) and Preissler and Carey (2005)), (ii) enhanced salience of the facial direction of the speaker, and (iii) usage of several different carrier phrases for the novel label ("That's a *peri.*", "See, it's a *peri.*", and "Look, it's a *peri.*").

Luyster and Lord (2009) found that only very few children in their ASD group made the error of incorrectly mapping the novel label to their own toy in the *discrepant labelling* condition, and that the proportion of children who did so in this group was not signicantly different from the proportion of children who did so in a TD control group. However, the children in the ASD group did not show clear signs of having mapped the novel label to the experimenter's toy either (i.e. they did not select the experimenter's toy more often than expected by chance in this condition, nor was there a significant difference between how often they selected the experimenter's toy in the *discrepant* compared to the *follow-in* condition). In sum, these results show that given the extra contextual supports that Luyster and Lord built in, a subsample of children with ASD were able to block an incorrect word-object mapping when the speaker's gaze was not directed at the object of the child's focus during labelling, even if they are

not able to then use the speaker's gaze to infer the correct referent. A similar pattern of results was reported by Baldwin (1991, 1993b) for TD infants of 16-17 months old (as opposed to infants of 18-19 months), indicating that this pattern corresponds to a stage of typical word learning development.

In line with the results of Luyster and Lord (2009), Parish-Morris et al. (2007) showed in their Experiment 2 that children with ASD can learn new words equally well as language-age matched and mental-age matched TD controls by observing the experimenter either pointing to or touching the target object. However, the ASD group showed reliable word learning only when the target object was interesting to them, in contrast to the two TD groups who showed reliable word learning both for objects they deemed interesting and objects they did not. Furthermore, Parish-Morris et al.'s Experiment 4 showed that children with ASD have difficulty learning words when intention-reading is required to infer the referent. Children in this experiment were forced to infer the experimenter's intended referent on the basis of her previous behaviour, in the absence of any observable social cues like eye gaze or pointing. Using an experimental design modelled on that of Tomasello and Barton (1994) and Akhtar and Tomasello (1996) (see Section 3.1.1), the experimenter in this condition told the child she was looking for the *parlu* (novel label) in a sack purse containing four novel and two familiar objects. In response to the first three novel object she extracted, the experimenter shook her head, frowned and said "That's not the *parlu*.", and put them back. She then pretended not to be able to find the fourth novel object (the *parlu*) and handed the purse to the child asking "Can you find the *parlu*?". The child's response was then coded for whether they selected the target novel object, one of the three nontarget novel objects, or one of the two familiar objects (for which they already knew the words prior to the experiment). Results showed that both the language-age and mental-age matched control groups selected the target object significantly more often than chance, whereas the ASD group did not.

In sum, the word learning experiments reviewed in this section show that on a group level, autistic children show difficulties with using social cues and intention-reading to guide word learning, but that there are significant individual differences. In line with these findings, longitudinal studies of children with ASD have shown that the amount of joint attention that autistic children engage in at an early age is a good predictor for their language outcome later on (Anderson et al., 2007; Siller and Sigman, 2008; Toth

et al., 2006)[3]. Similarly, a follow-up study by Kasari et al. (2012) with autistic children five years after they had received early intervention for improving joint attention skills in a randomized controlled trial (starting around age 3), showed that children in the intervention group had improved significantly more in their expressive vocabulary than children in the control group.

To recap, the empirical evidence reviewed in this section shows that mindreading abilities are important for language development. More specifically, this section focused on the importance of children's ability to infer speakers' communicative intentions for the process vocabulary acquisition. This is attested by word learning experiments run with typically developing children, as well as the language development and word learning of children with autism spectrum disorder.

3.2 The role of language in mindreading development

The term *mindreading* refers to a complex suite of skills that develops in several stages (see e.g. Apperly, 2011; Kovács et al., 2010; Wellman, 2014). Infants first learn, by the age of 6.5 months, that others' actions are goal-directed (e.g. Csibra, 2008). They then learn that these goals are likely to be driven by 'dispositions', such as preferences and perceptual access, which they understand around the first year of life (e.g. Luo and Beck, 2010). Thirdly, in some situations understanding goals and dispositions is not enough to explain or predict others' behaviour, and an understanding of knowledge and ignorance is necessary as well, which also comes online around the first year of life (e.g. Luo and Beck, 2010). Finally, children have to learn that, in addition to simple ignorance, people can have beliefs about the world that are false. This requires an understanding that others' representations of the world do not necessarily correspond to reality. As discussed in Chapter 2, such *false-belief understanding* is often viewed as a hallmark feature of a fully developed mindreading ability, and false-belief tasks are therefore often used as a litmus test for full-blown mindreading.

The classic paradigm for testing false-belief understanding is the so-called 'Sally-Anne task' (Wimmer and Perner, 1983; Baron-Cohen et al., 1985). In this task, children are presented with a story in which, for example, Sally first places a marble in a basket

[3]Joint attention is the sharing of attention between two individuals toward a third object, individual or event. That is, a shared awareness between two individuals that they are together experiencing the same thing at the same time (Tomasello and Carpenter, 2007). As such, join attention is one of the earliest indicators of infants' realisation that others might be attending to or thinking about something different from their own focus of attention.

and then leaves the room, after which Anne moves the marble from the basket to a box while Sally is away. Sally then returns and the child is asked where Sally will look for her marble. When this task uses an *explicit* elicited-response measure (e.g. pointing, answering verbally, or selecting one of two pictures to continue the story in a 'story card' version of the task), typically developing children usually start passing the task (i.e. answering that Sally will look in the basket rather than answering she will look in the box) around the age of four (see Wellman et al., 2001, for a review). The traditional take on false belief understanding is therefore that around this age a critical development happens where children start understanding that others' mental states are not direct reflections of reality, but rather *representations of* this reality, which are not necessarily accurate (e.g. Flavell et al., 1990).

However, research with *implicit* spontaneous-response measures using eye-tracking, have found evidence that infants can track false beliefs as early as 15 months old (e.g. Onishi and Baillargeon, 2005; Southgate et al., 2007). These findings suggest that the fact that children under the age of four typically fail elicited-response versions of the task may be due to other task demands, such as inhibition, executive control, problem solving or pragmatic understanding (Onishi and Baillargeon, 2005; Kovács et al., 2010; Rubio-Fernández and Geurts, 2012). In contrast, dual-process accounts explain these findings by positing that implicit and explicit mindreading result from two separate cognitive subsystems, which have different developmental trajectories.

According to the *two-systems* account of Apperly and Butterfill (2009) (see also Butterfill and Apperly, 2013), the implicit mindreading abilities we find in infants (and nonhuman animals, as discussed in Chapter 2) are the result of a capacity for tracking belief-like states which is cognitively efficient but inflexible. In contrast, the explicit mindreading abilities we find in older humans are, as Apperly and Butterfill argue, the result of later-developing processes which are more flexible but also more cognitively demanding, and likely depend on development of executive functions and language. A similar but subtly different interpretation is the *submentalizing* account of Heyes (2014b,a, 2015, 2018). This account shares the view that implicit mindreading is a product of processes that are fast, automatic, parallel, and use information that derives from genetic inheritance (also known as 'System 1' processes; Heyes, 2018, chapters 3 and 7), while explicit mindreading is a product of processes that are slow, effortful and serial, based partly on output from System 1 and partly on information that these processes generate themselves (also known as 'System 2' processes; Heyes, 2018,

chapters 3 and 7). However, Heyes' (2014b; 2014a; 2015; 2018) submentalizing account diverges from the two-systems account in that the latter hypothesises that implicit mindreading is a result of cognitive processes that are specialised for the representation of mental states (Butterfill and Apperly, 2013), while the former states that it is a result of domain-general mechanisms such as attention, learning and memory. The hypothesis that implicit and explicit mindreading depend on separate mechanisms is supported by the findings of Senju et al. (2009), who showed that adults with Asperger Syndrome (a milder, high-functioning variant of ASD) show the opposite pattern of results to typically developing infants. That is, Senju et al.'s Asperger Syndrome group passed an explicit false belief task with ease, while failing to track false beliefs spontaneously in an eye-tracking task that was passed by typically developing infants.

Aside from these different theoretical accounts, the evidence of false belief understanding in infants is currently under debate after a number of publication have shown a failure to replicate these findings (see Schuwerk et al., 2018; Powell et al., 2018, and the rest of the special issue of *Cognitive Development* titled "Understanding theory of mind in infancy and toddlerhood", vol. 46, 2018). Although current initiatives to address these discrepancies, such as the ManyBabies project 2 (Frank et al., 2018), look promising, clarity is yet to be reached. In sections 3.2.1 and 3.2.2 below, I will therefore focus on studies that used elicited-response measures rather than implicit measures. These sections will first discuss the role language plays in the mindreading development of typically developing children (Section 3.2.1), followed by a review of mindreading development in deaf individuals with language delays (3.2.2).

3.2.1 The role of language in the development of mindreading in typically developing children

There is a large body of research indicating that the development of children's language and mindreading abilities go hand-in-hand (see e.g. Astington and Baird, 2005). Milligan et al. (2007) present a meta-analysis of 104 such studies (with a combined sample size of 8,891 children) which used standardised measures of language ability and a false-belief task to assess mindreading ability. This meta-analysis showed that the correlation between these two skills is robust with a moderate to strong effect size, also when controlling for age as a mediating factor. Furthermore, Milligan et al. assessed the directionality of the relationship and showed that the largest average effect

size was for early language abilities predicting later false-belief task performance. The predictive relationship between early false-belief task performance and later language ability was also significant, but the effect in this direction was significantly less strong than the predictive relationship in the other direction.

Another group of studies focused on the extent to which children being exposed to conversation about mental states (e.g. through parents and siblings) plays a role in their mindreading development. These studies are reviewed by Slaughter and Peterson (2011), who divide them up into three types. Firstly, correlational studies have shown that there is a relationship between the extent to which children are exposed to conversation about mental states in their homes and their performance on various mindreading tasks. However, these studies cannot address the directionality of the effect. It could be that conversation about mental states promotes the development of mindreading skills, but it could equally be the case that children with advanced mindreading abilities prompt their family members to talk more about mental states. The directionality of the effect can be addressed by the second and third types of study: longitudinal and training studies. An example of each will be discussed below (for a more complete review of both types of studies, see Slaughter and Peterson, 2011; and for a meta-analysis of training studies of mindreading, see Hofmann et al., 2016).

Firstly, Taumoepeau and Ruffman (2006, 2008) report two longitudinal studies showing that exposure to mental state language at an early age predicts children's later performance on mindreading tasks. Taumoepeau and Ruffman (2006) measured the amount of mental state terms mothers used when describing pictures to their 15-month-old infants, as well as the size of the children's mental state vocabulary (based on parental reports). When the children were 24 months old, Taumoepeau and Ruffman (2006) measured these same two variables again, and additionally tested the children's performance on two emotional understanding tasks. Results showed that the amount of desire terms that mothers used when the infants were 15 months old predicted both the children's mental state vocabulary and their performance on one of the emotional understanding tasks at 24 months old. This predictive effect was found while controlling for socio-economic status, children's language ability at 15 months, and the mothers' own performance on two emotion recognition tasks. In contrast, Taumoepeau and Ruffman (2006) found no predictive relationship in the other direction: between children's mental state vocabulary at 15 months and the amount of desire language their mothers used at 24 months.

In a follow-up study, Taumoepeau and Ruffman (2008) tested the same participants again another nine months later (when the children were 33 months old). They found that at this age, mothers' conversation about other's thoughts and knowledge (rather than about desires) as measured at 24 months predicted both children's mental state vocabulary and their performance on one of the emotional understanding tasks at 33 months old (accounting for 11% of variance in children's emotional understanding). (Again this predictive effect was found after partialling out socio-economic status, children's language ability at 24 months, and the mothers' own emotion recognition proficiency.) Similarly to Taumoepeau and Ruffman (2006), no predictive relationship was found in the other direction: from children's mental state talk at 24 months to their mothers' mental state talk at 33 months.

Taken together, the studies of Taumoepeau and Ruffman (2006, 2008) suggest that parents scaffold children's mindreading development through language, starting with conversation about desires, and later moving on to conversation about thoughts and knowledge. Taumoepeau and Ruffman's findings show that it was this parental mental state talk which predicted children's later mental state vocabulary and understanding of emotions, not vice versa. This is consistent with the findings of Hughes et al. (2005), who showed in a large-scale longitudinal twin study that environmental (rather than genetic) factors explained the majority of variance in participants' performance on mindreading tasks. However, these longitudinal studies do not address the question *what aspect* of being exposed to conversation about mental states aids mindreading development. Several hypotheses have been put forward in this regard.

Firstly, language may be beneficial for mindreading development because it provides *labels* for mental states, in the form of mental state terms such as "believe", "think", "know", etc. (Olson, 1988). Secondly, language may be beneficial because it provides children with the necessary *representational format* to be able to build advanced mental state representations (especially representations of false belief). Specifically, this hypothesis proposes that the acquisition of *complement clauses*, such as "Sally thinks the marble is in the basket", provides this representational structure (de Villiers and Pyers, 2002). de Villiers and Pyers propose that the realisation of an open truth value at the position of the embedded proposition (the fact that when Sally thinks the marble is in the basket this is not necessarily true) plays an important role in the development from 'implicit' to 'explicit' mindreading (see also de Villiers, 2007; de Villiers and de Villiers, 2012). Thirdly, language acquisition might aid mindreading development not by means

of any structural aspect of language, but because *discourse* with other people confronts a child with the fact that others can have other perspectives, beliefs and knowledge than herself (Harris, 1996). What is special about language according to this hypothesis is that it provides a lot of explicit evidence of this fact (e.g. through questions, clarification requests or dissent). If such evidence would have to be acquired purely through non-linguistic interactions, i.e. inferred from observable behaviour of others, it would have a more implicit form and would presumably also be more difficult to obtain and less frequently available. These three hypotheses are of course not mutually exclusive; several of these aspects of language could have additive effects on mindreading development.

The second and third of the three hypotheses discussed above were tested explicitly in a training study by Lohmann and Tomasello (2003). In all conditions in this study, children interacted with an experimenter about different deceptive objects, which appeared to be one thing at first sight (e.g. a flower) but on closer inspection turned out to have another function (e.g. a pen). Children were also asked about the beliefs of a third person character (in the form of a hand puppet) about the deceptive objects. In the first, *full training* condition, the experimenter demonstrated the deceptive aspect of the objects using both mental state verbs ("think" and "know") and sentential complement constructions. Children were first shown the object and asked what they thought it was. They were then given the object to discover its 'real' function, after which they were asked to recall their previous belief as well as their current knowledge about the object. This was then summarised by the experimenter. Subsequently, children watched the hand puppet go through the same process of discovery, again narrated by the experimenter. Finally, they were asked about the puppet's new belief about the object.

In the second, *discourse only* condition, the experimenter demonstrated the deceptive aspect of the object through discourse but without using any mental state verbs or sentential complement constructions. For example, instead of being asked "What do you think this is?" children were asked, "What is this?", and instead of the experimenter saying "You thought it was a flower", she simply said "A flower.". In the third, *no language* condition, the experimenter attracted the child's attention to the two different guises of the object with very minimal use of language, using only the appropriate nonverbal emotional expressions combined with short verbal attention getters such as "Look!" and (while drawing the child's attention to the real function of

the object) "But now look!". Finally, the fourth, *sentential complement only* condition did not involve any highlighting of the deceptive nature of the object. Instead the experimenter simply talked about them as normal objects and asked the child about their attributes, while making sure to use mental state verbs and sentential complement constructions. Children in this condition were also encouraged to use sentential complements themselves, by asking them to summarise things the hand puppet had said that would require using such a construction. The same number of mental state verbs and sentential complements was used in this condition as in the *full training* condition. Children were tested on a false-belief task both before and after receiving one of these four types of training (over the course of four separate sessions).

Lohmann and Tomasello's (2003) results showed firstly that language *is* a necessary precondition for children to benefit from false belief training; the children in the *no language* condition did not perform any better after training than before training. Secondly, improvement on the false belief task was found both in the *discourse only* and in the *sentential complement only* conditions, indicating that both talk about mental states and the specific representational structure of sentential embedding contribute to the benefit that language provides for mindreading development. The *sentential complement only* condition did involve mental state verbs as well though, which could also explain the positive effect of training. However, a comparison was made between two versions of the *full training* condition: one in which mental state verbs ("think" and "know") were used, and one in which only a communication verb ("say") was used, and this comparison revealed no difference. This last finding can be interpreted as at least indirect evidence that the main effect in the *sentential complement only* condition cannot have depended on the use of labels for mental states. Finally, the strongest improvement on the false belief task was found in the *full training* condition, suggesting that the contributions that mental state discourse and sentential complement constructions make to false belief understanding are independent from each other (because they add up in the *full training* condition). (Note that the *full training* condition did not involve more overall talk than the other conditions.)

In sum, Lohmann and Tomasello's (2003) training study demonstrates a clear causal link between language and mindreading development, where exposure to certain aspects of language speeds up the development of false belief understanding. However, the studies reviewed in this section cannot tell us whether exposure to language merely speeds up mindreading development, or whether it is also a *necessary* prerequisite. This

latter question is addressed by the case of mindreading development in deaf individuals with delayed language exposure, which is discussed in the next section.

3.2.2 Mindreading development in deaf children

It has been suggested that the case of deaf children with language delays in a sense constitutes the 'mirror image' of children with autism (Schick et al., 2007). That is, children with high-functioning autism, despite having difficulty engaging in social interactions, are able to employ mental state language in order to achieve explicit reasoning about false beliefs (Tager-Flusberg and Joseph, 2005; Senju et al., 2009). Deaf children with a delayed exposure to language on the other hand are amply endowed with interest in social behaviour, but may be limited in developing full-blown mindreading because of a lack of access to the aspects of language that aid this type of reasoning. The current section will review evidence for the latter claim.

Mindreading development after delayed exposure to an established sign language

Evidence in favour of the hypothesis that access to language is important for developing full-blown mindreading comes from studies of deaf children of hearing parents, who do not enter a community of fluent signers until they start primary school. Peterson and Siegal (2000), Meristo et al. (2011) and Pyers and de Villiers (2013) each review such studies (15 in total), together covering a wide range of different family circumstances, approaches to deaf education, and ways of measuring mindreading ability. These studies consistently show that the mindreading development of deaf children from hearing families is delayed compared to that of both hearing children and deaf native signers (i.e. deaf children growing up with deaf, signing parents). In contrast, deaf native signers develop their mindreading abilities around the same age as hearing children. Peterson and Siegal (2000), Meristo et al. (2011) and Pyers and de Villiers (2013) each interpret these findings as evidence that early conversational interactions are important for mindreading development (see also Peterson and Siegal, 1995). I will discuss some examples of these studies below.

Schick et al. (2007) measured both mindreading ability and several aspects of language ability in a study comparing deaf children of hearing parents, deaf native signers, and hearing children of hearing parents (all between 4 and 8 years old). Schick et al.

found that both hearing children and deaf native signers outperformed deaf children of hearing parents on various tests of mindreading ability (even when the verbal demands of the task were minimised). Furthermore, Schick et al. found that deaf-of-hearing children's performance on the false belief tasks was predicted by their ability to comprehend sentential complement sentences (and, in the case of verbal false belief tasks, also by their receptive vocabulary). However, the extent to which mental state talk was present in the language *input* for these children was not measured in this study.

More direct evidence for the hypothesis that a lack of mental state conversation contributes to the delayed mindreading development found in deaf children of hearing parents, comes from two studies showing that even when hearing parents of deaf children make great effort to learn sign language in order to communicate with their deaf child, their conversations remain impoverished compared to those of hearing children and deaf native signers with their parents. Moeller and Schick (2006) measured the amount of mental state terms uttered by hearing mothers of deaf children compared to hearing mothers of hearing children. Moeller and Schick used three different tasks to elicit mental state conversation: a toy-play session, a session of talking about family photos, and a session of watching and discussing short silent movies of characters reacting to unexpected events. Moeller and Schick also measured the children's scores on several different false belief tasks.

In line with the findings summarised above, Moeller and Schick (2006) found that although the hearing children in their sample were younger than the deaf children (mean age 5.0 versus 6.9 respectively), they outperformed the deaf children on two false belief tasks. Furthermore, Moeller and Schick found that the hearing mothers of deaf children in their sample used significantly fewer and less diverse mental state terms in their conversations with their children than did the hearing mothers of hearing children. Finally, regression analyses showed that maternal mental state talk accounted for a significant amount of variance in the children's scores on the false belief tasks (as opposed to 'non-mental state' maternal language input). However, as Moeller and Schick note themselves, it is hard to determine the direction of causality based on these correlational findings; it could be that hearing mothers of deaf children simply learn sign language as required to keep up with their child's development, such that they start discussing mental states when their child starts showing an understanding of these concepts. This alternative explanation could be addressed with longitudinal research like that of Taumoepeau and Ruffman (2006, 2008), discussed in Section 3.2.1.

Morgan et al. (2014) found similar results for 2-3 year old deaf children with cochlear implants (who thus have some limited hearing abilities). Morgan et al. elicited mental state conversation by asking mothers to talk with their children about pictures of emotionally charged or 'mentalistic' situations (using any or all forms of spoken, signed and gestural communication). Morgan et al. characterised the mental state talk that was used in the hearing-deaf dyads as similar to the input received by hearing children of a younger age, to whom parents talk more about desires and less about epistemic mental states like "think" and "know" (see discussion of Taumoepeau and Ruffman, 2006, 2008, in Section 3.2.1). Based on these findings, Morgan et al. speculate that hearing parents of deaf children adapt their mental state talk to fit the perceived level of understanding of their child, leading to simpler conversations. Unfortunately, this study did not include any measure of mindreading development of the children.

Finally, O'Reilly et al. (2014) showed that some differences in mindreading ability between late signers compared to deaf native signers and hearing individuals are still detectable in adulthood. Using more advanced mindreading tasks (second-order false belief understanding) and several tests of conversational sarcasm understanding, O'Reilly et al. found that late signers around the age of 40 were still outperformed by both native signers and hearing participants of the same age, despite decades of social experience as a fluent signer. These protracted effects of delayed exposure to language are even more dramatic when the sign language of the community that deaf people are born into is only just starting to develop. This case will be discussed in the next section.

Mindreading development in the absence of an established sign language

More dramatic delays of mindreading development were found in a study of signers of the emerging Nicaraguan Sign Language (NSL), by Pyers and Senghas (2009). NSL began to develop in the 1970s when deaf children in Nicaragua started to enter special-education schools. Younger children who entered these schools since have acquired NSL from their older peers, and continued developing the language. Research on NSL often divides its users up into ten-year cohorts. This is an artificial divide given that the flow of children entering the language community is continuous, but it allows researchers to investigate groups of users roughly according to how developed the language was when they first acquired it.

Pyers and Senghas assessed the performance of the first and second of these user cohorts on a minimally linguistic false-belief task across two time points. The first assessment ('Time 1' below) took place when cohort 1 had a mean age of 27 and cohort 2 a mean age of 17, and the second assessment ('Time 2') took place two years later (with mean ages 29 and 19). Pyers and Senghas found that at Time 1, the second cohort outperformed the first cohort on the false belief task, even though both cohorts had started learning their language around the same age, and the first cohort had ten more years of general social experience. Instead, the crucial difference between these two cohorts seems to be their language experience: as the pioneers of NSL, the first cohort had acquired a somewhat less-developed version of the language than the second cohort. Pyers and Senghas measured specifically how many mental state verbs the signers produced in an elicitation task, and found that at Time 1, the first-cohort participants signed significantly less tokens of mental-state verbs than the second-cohort participants did. In fact, half of the first-cohort participants did not produce any mental state verbs at all. An analysis of individual differences revealed that all first-cohort participants who had not produced any mental state verbs, also failed the false belief task.

At Time 2, two years later, there was no longer a difference between the two cohorts in terms of how many mental-state verbs they produced on the same elicitation task. Post-hoc analysis showed that the disappearance of this difference was due to the first-cohort signers having increased their use of mental-state verbs (every participant now produced at least one). The first-cohort signers also improved significantly in their performance on the false belief task from Time 1 to Time 2, leading to the disappearance of a significant difference in false belief task performance between the two cohorts. The false belief task used by Pyers and Senghas (2009) is simpler in terms of the mental state reasoning it requires than the mindreading tasks used in the study of O'Reilly et al. (2014) which was discussed in the previous section. O'Reilly et al. (2014) tested participants' understanding of second-order false beliefs and conversational sarcasm, which both come online much later in typical development than first-order false belief understanding (around the age of 10, see Perner and Wimmer, 1985 for second-order false beliefs, and Peterson et al., 2012 for sarcasm). As discussed above, O'Reilly et al. (2014) found that when tested on these tasks, even delayed-exposure signers who had learned an established sign language were still outperformed by both native signers and hearing participants around the age of 40. Thus, although having caught up in terms

of first-order false belief understanding, the first-cohort signers might still have shown difficulties on such harder mindreading tasks if those had been included in the Pyers and Senghas study.

Pyers and Senghas (2009) also remark that between their Time 1 and Time 2 assessments, newly adult second-cohort NSL signers started socialising at a deaf association, thereby increasing the contact between first- and second-cohort signers. This may well have caused the expansion of the first cohort's mental state vocabulary. Pyers and Senghas interpreted these findings as evidence that language plays a crucial role in the development of false belief understanding. As they note themselves, their findings are compatible with any of the three hypotheses about the role of language in mindreading development discussed in Section 3.2.1 (labels, sentential complements, and increased exposure to diverging perspectives through discourse). However, the findings are not compatible with the hypothesis that explicit false-belief understanding can be acquired through social experience alone, without exposure to language.

Following up on this study, Gagne and Coppola (2017) asked what would happen to mindreading development if individuals are *never* exposed to a conventional language. This is the case for deaf people who grow up in isolation from a deaf community (due to an absence of deaf schooling). These individuals usually develop a gestural communication system which they use to communicate with family members, known as *homesign*. Gagne and Coppola tested a group of homesigners in Nicaragua on several social cognition tasks, tapping into different stages of typical mindreading development, from level-1 visual perspective-taking (i.e. an understanding of what another person can and cannot see) through to first-order false-belief understanding (using a task that minimised the need for communication and maximised the role of first-hand visual experiences). Gagne and Coppola then compared the results of the homesigners to those of participants from the first cohort of NSL signers, and hearing speakers of Spanish with minimal schooling (to match the homesigners' level of education).

Gagne and Coppola (2017) found that the performance of the homesigners did not differ from that of the two comparison groups on tasks measuring visual perspective-taking (levels 1 and 2) and a task of photographic misrepresentation (testing understanding of the fact that the current state of the world does not necessarily reflect a previous state). However, the homesigners were outperformed by both comparison groups on the false-belief task. These results suggest that whereas visual perspective-taking and photographic misrepresentation can be learned through decades of socio-

visual experience (the homesigners in this study were between 26 and 35 years old), false-belief understanding is dependent on exposure to a conventional language shared with a community of speakers. The development of an idiosyncratic communication system like homesign is apparently not sufficient to unlock this development.

To recap, the empirical evidence reviewed in this section shows that exposure to language is important for the development of full-blown mindreading. This is attested both by individual differences in mindreading being predicted by language exposure in typical development, and by mindreading development being delayed when exposure to language comes late.

3.3 Computational models of word learning

As reviewed in Section 3.1, the ability to infer a speaker's communicative intentions is important for developing typical language and is used by children in the process of word learning. However, computational models designed to account for children's impressive word learning abilities have only recently started to incorporate such intention-reading. The challenge that such models have sought to address is that of *referential uncertainty*. This refers to the problem that words are used in complex environments, and that when a learner is confronted with a novel word, it could in principle label any part of that complex environment. Even worse, the word could refer to an object or event which is not currently perceivable to the learner, because it labels something that is spatially or temporally distant from the context in which it is uttered, or because it describes an abstract concept. Therefore, every time a word is used, there are many possible meanings which the learner could infer as the word's intended meaning. Such referential uncertainty can in theory be unbounded: if all logically possible meanings of any novel word are equally plausible candidates in any context of use, a learner would never be able to infer the meaning of a novel word (an observation often ascribed to Quine, 1960, in his work on radical translation). Thus, a word learner needs to have ways to reduce the hypothesis space of possible meanings considerably.

Several different types of solutions to this problem have been put forward and tested using mathematical and computational models. These can be roughly divided into three kinds: (i) solutions based on learning biases, (ii) solutions using social cues, and (iii) solutions based on intention-reading. Each of these solutions builds on an underlying mechanism of associative learning which simply strengthens associations

between a word and all the meanings that are present in the context in which it is uttered. When applied over a sufficient amount of different contexts, such associative learning can eventually converge on the word's 'true' meaning through a mechanism known as cross-situational learning (Siskind, 1996). Cross-situational learning is based on the principle that words occur in different contexts, and that candidate meanings can progressively be ruled out on the basis that they co-occurred with the word of interest in one context but not another. Thus, by observing a sufficient amount of different co-occurrences between a given word and the objects or events in the context of which it is uttered, a learner should gradually be able to narrow down the set of candidate meanings to the true meaning of the word (see Akhtar and Montague, 1999 and Smith and Yu, 2008 for empirical evidence of children using this principle in word learning).

Using mathematical modelling, Blythe et al. (2010, 2016) showed that if such cross-situational learning is combined with heuristics for reducing the hypothesis space of possible meanings (which could be instantiated by any or all of the three mechanisms mentioned above: learning biases, social cue-following or intention-reading), this can lead to powerful vocabulary-learning abilities. Moreover, they found this to be the case even if the heuristics themselves are weak and unable to eliminate all referential uncertainty. Specifically, Blythe et al. (2010) showed that when cross-situational learning is combined with heuristics that constrain the set of candidate meanings, this allows for the learning of large lexicons, even under high levels of referential uncertainty. Building on these findings, Blythe et al. (2016) showed that even infinite referential uncertainty can be overcome using this model, as long as the learner is able to rank candidate meanings in terms of their plausibility. Models that have implemented more specific instantiations of the three different solutions for reducing the space of possible meanings mentioned above (learning biases, social cue-following, and intention-reading) are discussed in turn below.

3.3.1 Solutions using learning biases

Children have been shown to place several constraints on possible meanings in word learning, which helps reduce the problem of referential uncertainty. For instance, Markman (1990) reviews empirical evidence showing that children use the *whole object assumption*: an expectation that novel words will refer to an object as a whole, rather

than one of its parts or other properties. This assumption comes with the disadvantage that languages contain many words that do in fact refer to parts or other properties of objects. Markman proposes that one of the ways in which children overcome this problem is by counterbalancing their whole object assumption with another constraint: the *mutual exclusivity* principle. The latter bias was first empirically demonstrated in children's word learning by Markman and Wachtel (1988), and consists of the assumption that a novel word will not refer to something that already has a label (i.e. something that the child already knows the word for). However, this heuristic comes at the cost of making it harder to learn synonyms (Liittschwager and Markman, 1994). Finally, children have also been shown to use the syntactic context and argument structure in which words occur in order to constrain the set of possible meanings; a phenomenon known as *syntactic bootstrapping* (Gleitman, 1990; see also Naigles and Swensen, 2008, for a review of empirical evidence).

Kachergis et al. (2012) implemented a combination of cross-situational learning and the mutual exclusivity principle in an associative model through a trade-off between two very simple attention mechanisms. Firstly, the learner's attention is directed towards word-object pairs that have already been encountered previously, but secondly, more attention is also drawn to words and objects that do not yet have an association (i.e. novel items). The first attention bias implements cross-situational learning, where those word-object pairings that occur across a set of different contexts will be reinforced more strongly than others. The trade-off between the familiarity and novelty attention biases however gives rise to mutual exclusivity effects: if a novel word co-occurs with a novel object, this pairing will be reinforced more strongly than the potential pairing of the novel word with a familiar, already-labelled object.

3.3.2 Solutions using social cues

The word learning strategies described above can be used independently of making any observations about the speaker uttering the words. However, actual word learning does not happen in such a social void. As reviewed in Section 3.1, empirical research has shown that children are sensitive to social cues like eye gaze, pointing and joint attention during word learning (see also Paulus and Fikkert, 2014; Yu and Smith, 2012). Yu and Ballard (2007) integrated such social cues in a computational model of statistical word learning. As training and testing data for this model, Yu and Ballard used videos of

naturalistic mother-child interactions from the CHILDES corpus (MacWhinney, 2000). These videos were given as input to the learning model in two 'streams': a transcript of the words that were uttered, and a description of which objects were in view at the moment of utterance.

In addition to this basic input, Yu and Ballard (2007) also transcribed the social cues that were provided by the mothers in the form of prosody and joint attention. Specifically, they used acoustic measures to determine which words in the speech stream received prosodic emphasis, and gave those words additional weight in the associative learning algorithm. Similarly, Yu and Ballard encoded whether objects in the video were in joint attention between mother and child, and weighted these referents accordingly. Using the standardised measures of *precision* (the proportion of learned word-object pairings that were correct) and *recall* (the proportion of total correct pairings that were found by the model), Yu and Ballard found that the integrated model using both prosodic and joint attention cues performed better than (i) the baseline statistical learning model, (ii) a model using only prosodic cues, and (iii) a model using only joint attention cues.

3.3.3 Solutions using intention-reading

Finally, as reviewed in Section 3.1, children do not only use directly observable social cues which direct attention to certain words or objects (like prosody and joint attention), but also make inferences about *un*observable referential intentions to guide their word learning. Frank et al. (2009b) incorporated such intention-reading in a rational (Bayesian) model of cross-situational word learning. Instead of re-weighting words and referents based on social cues, Frank et al. assumed that the learner posits an unobserved variable which mediates between the objects present in the physical environment and the words that the speaker utters. This unobservable variable is a model of the speaker's referential intention.

Similar to the input that Yu and Ballard (2007) gave to their baseline statistical learning model, Frank et al.'s (2009b) learner observes the combinations of words uttered and objects present in the context at the moment of utterance. Based on these word+context observations, the Bayesian learner then evaluates every logically possible lexicon based on how likely the observed word is given that lexicon combined with every possible referential intention. The full set of possible referential intentions that

the learner considers simply consists of all possible subsets of the objects present in the context (including the possibility of an 'empty' intention). The resulting likelihood of the observed data given a lexicon hypothesis and all possible referential intentions is then combined with a prior that favours parsimonious lexicons, together yielding a posterior probability distribution over all possible lexicons. Thus, lexicons that associate words with objects that they do not co-occur with will be considered less probable than lexicons that map words to objects they do co-occur with; this model therefore instantiates cross-situational learning. In addition however, Frank et al. allowed for words to be uttered 'nonreferentially', meaning they are used without a consistent referent. This possibility was added to enable the learner to deal with verbs, adjectives, function words and nouns referring to objects that were not present in the context.

Frank et al. (2009b) tested their model on its ability to learn concrete nouns from the same two CHILDES videos (MacWhinney, 2000) that Yu and Ballard (2007) used to test their model. For this purpose, Frank et al. selected the lexicon with maximum a posteriori probability given the observed corpus of contexts+utterances, and measured its precision and recall. In addition, Frank et al. measured the precision and recall of the model's inferences about the speaker's referential intentions, by choosing the intentions with the highest posterior probability given the best lexicon. Frank et al. found that on both these measures the intention-reading model outperformed several alternative associative learning models, including the baseline statistical learning model used by Yu and Ballard.

Frank et al. (2009b) attribute the high precision of their model compared to these other models (i.e. the fact that it learned relatively few incorrect mappings) to two factors. Firstly, the fact that the model can distinguish between words that are used referentially and words that are used nonreferentially allows it to leave words that were used without a consistent referent out of the lexicon. Secondly, the fact that the model considers empty intentions as well as referential intentions means that it can disregard utterances that do not refer to any of the present objects. Although Frank et al.'s implementation is rather simplified compared to what intention-reading amounts to in real-life vocabulary learning, their model clearly benefits from its ability to assume a mediating factor between the context and a speaker's utterances: the speaker's referential intention.

In sum, Yu and Ballard (2007) and Frank et al. (2009b) models make significant progress in incorporating social cues and intention-reading in computational models of

word learning, but they do not yet take into account the evidence reviewed in Section 3.2, showing that such intention-reading skills may in part be learned. Instead, these models treat the ability to make use of social cues and infer referential intentions as a given and fixed capacity, fully present from the very start of the vocabulary-learning process. As reviewed in sections 3.1 and 3.2 however, both the ability to learn words and the ability to infer intentions improve over developmental time, and this may progress partly by virtue of a co-development between these two skills. In Section 3.4, I present an agent-based model of word learning which takes these developmental dynamics into account. This model assumes a Bayesian learner who is faced with a joint-inference task: in addition to inferring a lexicon, they have to simultaneously infer the speaker's perspective on the world, which affects the speaker's referential intentions. In this model, learning about the speaker's lexicon and learning about their perspective are inextricably bound together.

3.4 An integrated model of perspective-taking in word learning

In this section I present a model in which agents simultaneously learn about a speaker's lexicon and about their (the speaker's) perspective on the world. Simulation results obtained with this model (see Section 3.5) show that learners can correctly infer both a speaker's lexicon and their perspective from scratch, by bootstrapping one from the other. However, this is only possible if (i) the learner is able to represent the speaker's perspective, and (ii) the speaker's lexicon has informative mappings from signals to referents (i.e. is not completely ambiguous). If either of these two conditions is not met, the learner will not be able to infer the correct combination of the speaker's lexicon and perspective. In this model, learning a lexicon and learning about a speaker's perspective are thus interdependent.

3.4.1 Mental States

As a simple model of how mental states influence communicative behaviour, I implement mental states as a probability distribution over potential referents, from which referential intentions are then sampled. The world consists of a fixed set of potential referents, which are visible to all agents, and each referent has a single attribute. The

values of these attributes can vary, and these variations create different contexts (i.e. states of the world). Each agent has an individual 'perspective' on these contexts (i.e. a particular view on the world), which is a parameter in a function that maps from the context to the agent's probability distribution over referents (i.e. their 'mental state'). In every communicative interaction, the speaker agent samples a referential intention from their probability distribution over referents given the current context (i.e. chooses a referential intention based on their current mental state), and produces a corresponding utterance based on their lexicon (see Section 3.4.2). In other words, in each interaction the current state of the world combined with the agent's view on the world render certain referents more salient to them than others, and this determines how likely they are to talk about each referent in that context.

In all simulations reported in this thesis, an agent's perspective remains stable throughout their lifetime; what changes with each interaction is the context in which agents communicate. This means that the probability distribution over referents will be different for two agents with two different perspectives on the same context, and also for the same agent in two different contexts. In all simulations reported in this thesis, the set of possible perspectives is limited to two, which are exactly opposite to each other (as depicted in Figure 3.1). An agent's perspective can be interpreted in a literal sense, where the physical proximity of objects influences their salience (as illustrated in Figure 3.1), but it can equally be interpreted in a more abstract sense; as a world view that determines what potential topic of conversation is most salient to the agent in a given context.

Importantly, all referents that exist in the world are always considered potential referents. That is, each referent's probability of being chosen as the speaker's referential intention is always nonzero. This means that no potential referent can be excluded on the basis of cross-situational learning, and referential uncertainty is infinite in this sense. The only way in which contexts differ from one another is in the combination of attributes of the different referents, and observing many different contexts is what gives the learner a way into learning about the speaker's perspective. To continue with the spatial interpretation of this model mentioned above, the probability that a given object will be chosen as referent in a given context is equal to the inverse of the distance between the agent's perspective and that object (i.e. the object's 'salience'), normalised over the full set of objects, as shown in Equation 3.1.

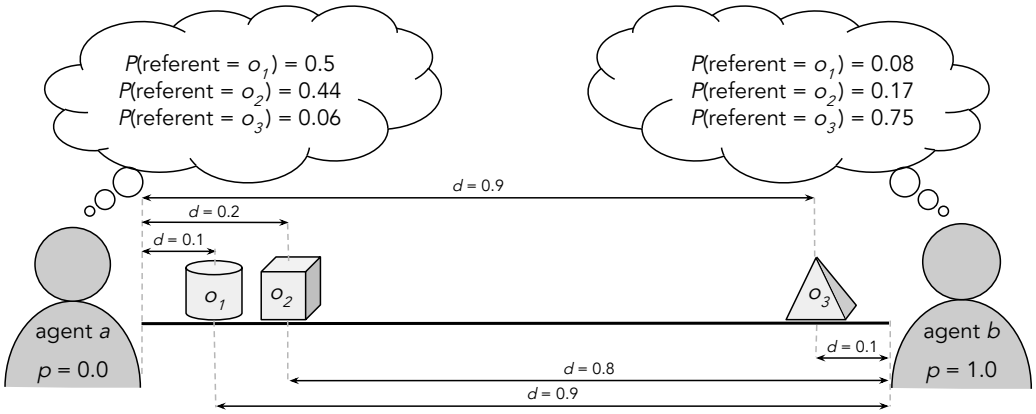


Figure 3.1: Diagram of how a context and an agent's perspective together give rise to a probability distribution over potential referents (illustrated using spatial interpretation of perspectives and referent attributes as described above). Here, the perspectives (p) of the two agents are diametrically opposed; hence their position on opposite extremities of the 'context' line. d stands for the distance between an agent and an object. Thus, object 1 (o_1) is equally close to agent a as object 3 (o_3) is to agent b. However, agent b is 1.5 times more likely to choose o_3 as an intended referent than agent a is to choose o_1 (see the agent's thought bubbles for the probabilities with which they will choose each of the objects as their referential intention). This asymmetry is a result of the distances (and therefore the saliencies) of the other objects in the context. For agent a, o_2 is also very salient, which means the probability mass over potential referents has to be distributed roughly equally over o_1 and o_2. For agent b in contrast, the next most salient object (o_2) is much further away than their most salient o_3, which means the ratios between the probabilities of choosing objects as intended referents work out differently. This difference in intention probability ratios between the two perspectives is a result of the way mental states are calculated from a set of saliency values (see Equation 3.1): the model assumes that saliencies are relative. Therefore, saliency values are normalised to yield a probability distribution over communicative intentions. Note that because of these differences in ratios between intention probabilities, a learner could potentially distinguish between these two perspectives even without knowing what signals map to which objects, as long as they get to observe enough different contexts, and the speaker's signals are not completely ambiguously mapped to their intended referents. However, this 'lexicon-independent' learning strategy will become less effective the more ambiguity the speaker's lexicon contains.

$$P(o_i = r|p, c) = \frac{1 - |p - o_{i_c}|}{\sum_{o \in O} 1 - |p - o_c|} \tag{3.1}$$

where r stands for referent, p stands for the perspective of the agent in question, o_c stands for the attribute of object o in context c (which we can think of as the object's spatial location), and O refers to the full set of objects. Because the saliency values are normalised in order to yield the probability distribution over referential intentions,

different combinations of saliency values will yield different probability distributions (see Figure 3.1 for a concrete example). Even if a learner does not know how the utterances of a speaker map to individual referents, they will still be able to evaluate the probability of different perspective hypotheses on the basis of ratio differences between the intention probabilities that those different perspective hypotheses predict. However, this learning strategy depends on two conditions: (i) that the learner gets to observe a sufficient amount of different contexts, and (ii) that the speaker does not use signals in a completely ambiguous way (where each signal could refer to each object). Below I explain how lexicons are implemented in this model.

3.4.2 Lexicons

Aside from a perspective, each agent has a lexicon which determines what signal(s) they use for a given referent. Lexicons consist of discrete binary mappings between signals and referents. The probability of a signal s being uttered for a referent r given lexicon ℓ is shown in Equation 3.2.

$$P(s \mid r, \ell) = \begin{cases} \dfrac{1-\epsilon}{|s_r|} \text{ if } s \text{ maps to } r \text{ in } \ell \\ \dfrac{\epsilon}{|\mathcal{S}| - |s_r|} \text{ otherwise} \end{cases} \tag{3.2}$$

where $|s_r|$ stands for the number of signals that map to referent r in lexicon ℓ, $|\mathcal{S}|$ stands for the total number of signals in ℓ, and ϵ stands for the probability of making an error in production (where an error is a random choice between the signals which are *not* associated with the intended referent). The probability of production errors ϵ was fixed at 0.05 throughout all simulations reported in this thesis.

The space of all logically possible lexicons expands exponentially as lexicon size increases (i.e. if referents or signals are added). For example, in a world where there are two referents and two signals, there are nine logically possible lexicons (referent 1 maps to signal *a*, signal *b*, or either, and referent 2 independently maps to signal *a*, signal *b*, or either). And by the same logic, a world with three referents and three signals yields 343 possible lexicons. Lexicons in which a referent does not have any signals associated with it are excluded, because if a given referent is associated with *none* of the signals or *all* of the signals, these both result in exactly equivalent production behaviour (namely $P(s) = 1/|R|$ for every s). The amount of possible lexicons is then given by $|L| = (2^{|S|} - 1)^{|R|}$, where $|L|$ stands for the total number of lexicons, $|S|$ for

the total number of signals, and $|R|$ for the total number of potential referents. (The base here is $2^{|S|} - 1$ because lexicons in which a referent has no signals associated with it are excluded.) All simulation results reported in this thesis were obtained with 3x3 lexicons.[4] However, see Woensdregt et al. (2016) for the developmental results obtained with the same model and 2x2 lexicons.

3.4.3 Learning

The task of the learner in this model is to simultaneously infer both the perspective and the lexicon of a speaker who provides input, based on observing what signals the speaker utters in different contexts. In other words, the learner has to infer two variables which are unobservable: the speaker's perspective and the speaker's lexicon, based on two variables which *are* observable: the speaker's utterances in context. Figure 3.2 depicts how these unobservable and observable variables are related. The world randomly creates a context which is observable, and that context combined with the speaker's perspective gives rise to the speaker's probability distribution over referents, from which a referential intention is sampled. That referential intention combined with the speaker's lexicon then leads to an utterance, which is observable to the learner.

The learner considers a hypothesis space consisting of all possible combinations of perspective hypotheses (of which there are two) and lexicon hypotheses (of which there are 343). For each of these composite hypotheses, the learner updates their belief in the hypothesis based on the data (observed combinations of context and utterance), using Bayesian inference. The posterior probability of a composite hypothesis given a set of data is proportional to the likelihood of the data given the composite hypothesis, times the prior probability of the composite hypothesis, as shown in Equation 3.3.

$$P(\ell, p \mid \mathcal{D}) \propto P(\mathcal{D} \mid \ell, p) P(\ell, p) \tag{3.3}$$

The likelihood of a dataset $\mathcal{D}$ given a composite hypothesis (ℓ, p) is given by Equation 3.4.

[4] A world in which there are only two objects regularly gives rise to symmetrical contexts in which the saliency distributions for the two possible perspectives are each other's exact mirror image. Combined with the fact that every lexicon also has a mirror image, this results in every input combination of speaker lexicon + perspective having one exceptionally strong competitor in the learning process; namely the hypothesis that the speaker has exactly the opposite perspective and the mirror image lexicon to what they really have. To prevent this somewhat artificial scenario from affecting simulation results, this thesis instead uses 3x3 lexicons to reduce the probability of the world throwing up symmetrical contexts.

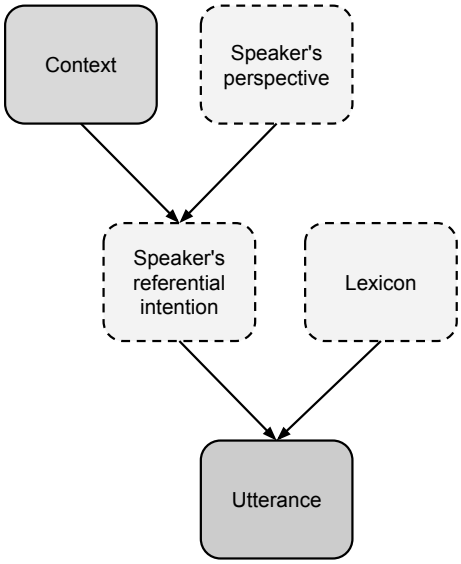


Figure 3.2: Diagram of how communicative behaviour is determined. Variables in dark grey are observable to the learner, variables in light grey are unobservable. The learner has to infer the perspective and lexicon of their cultural parent based on observations of the parent's utterances in context.

$$P(\mathcal{D} \mid \ell, p) = \prod_{d \in \mathcal{D}} P(s_d \mid \ell, p, c_d) \tag{3.4}$$

where each individual data point d consists of a context c_d and a signal s_d that was uttered by the speaker in that particular context. The likelihood of a single utterance s_d given a composite hypothesis and a context c_d is given by marginalising the product of the probability of the signal given the referent and the probability of the referent being intended over all objects in c_d, as shown in Equation 3.5.

$$P(s_d \mid \ell, p, c_d) = \sum_{o \in c_d} P(r_o \mid p, c_d) P(s_d \mid r_o, \ell) \tag{3.5}$$

where o stands for object and $P(r_o)$ for the probability of object o being chosen as the speaker's intended referent.

3.4.4 Priors

In all simulations reported in this chapter, the learner has a uniform prior over lexicons (i.e. presupposes that all lexicons are equally probable before seeing any data). For the learner's a priori expectations about perspectives however, two different priors

are compared below: a *neutral prior* which assumes both perspectives are equally likely, and an *egocentric prior* which assumes the speaker is more likely to share the learner's perspective than to have a different perspective. This egocentric prior assigns a probability of 0.9 to the hypothesis that the speaker's perspective is the same as the learner's (and thus a prior probability of 0.1 to the opposite perspective hypothesis). This egocentric bias is motivated by empirical evidence showing that young children start out reasoning about other minds from an egocentric perspective, and that this bias diminishes over time (see Birch and Bloom, 2004, for a review). Because this bias is rooted in empirical evidence, the egocentric learner case is of particular interest for the purposes of this thesis. Over all combinations of lexicon and perspective prior, the prior probability of a composite hypothesis (ℓ, p) is simply the product of the prior probabilities of the ℓ and p hypotheses separately, i.e. $P(\ell, p) = P(\ell)P(p)$.

3.5 Simulation results: co-development of lexicon-learning and perspective-inference

In this section, I will present simulation results obtained with the model described in Section 3.4, focusing on how learners with different prior biases deal with input from different types of speakers.[5] As described in Section 3.4.4 above, I compare two different types of learner: one with a uniform prior over both lexicon hypotheses and perspective hypotheses (henceforth *unbiased learner*), and one with a uniform prior over lexicons but a strong egocentric perspective bias (henceforth *egocentric learner*). In all figures below, the results for the unbiased learner are depicted on the left-hand side, and the results of the egocentric learner are depicted on the right-hand side. Firstly, this section will address the question under what circumstances these two learners can successfully learn the correct combination of perspective and lexicon hypothesis about a given speaker. Secondly, the results presented here will address the question how the speaker's lexicon affects the timecourse of this learning process.

In order to answer the second question, I categorise the set of all possible lexicons by their informativeness. Informativeness can be formalised in different ways, but for this thesis I chose an operationalisation based on communicative success, given that this is the measure which becomes most relevant in the upcoming chapters, in which

[5] All the code that was used to run these simulations is freely available at https://github.com/marieke-woensdregt.

populations of agents are subjected to different selection pressures, including one on the basis of communicative success. For the purposes of the current chapter, I measure the communicative success of a lexicon 'with itself' as a measure of its informativeness. This measure can however equally be thought of as quantifying the communicative success between two agents who share the same lexicon, and this is how it will be described below (to maximise continuity with the upcoming chapters).

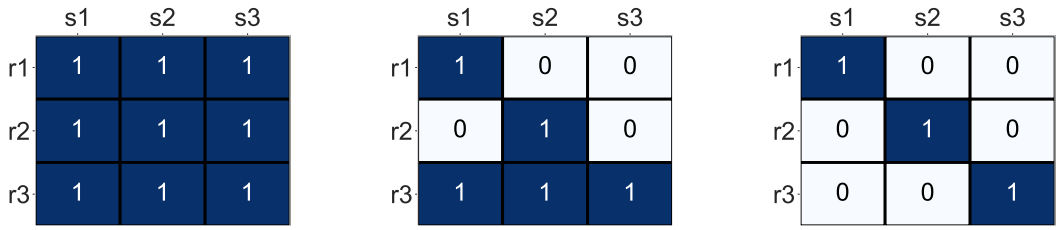
The communicative success CS between two agents i and j is measured as the mean of their communicative accuracy (ca) in both directions (i.e. when agent i is the speaker and agent j the hearer, and vice versa), as shown in Equation 3.6.

$$CS = \frac{ca(\ell_i, \ell_j) + ca(\ell_j, \ell_i)}{2} \tag{3.6}$$

where ℓ_i stands for the lexicon of agent i, ℓ_j for the lexicon of agent j, and $ca(\ell_i, \ell_j)$ for the communicative accuracy between a speaker with lexicon ℓ_i and a hearer with lexicon ℓ_j. $ca(\ell_i, \ell_j)$ in turn is defined as the average probability that speaker i will produce a signal which enables hearer j to correctly identify i's intended referent, as shown in Equation 3.7.

$$ca(\ell_i, \ell_j) = \frac{1}{|R|} \sum_{r \in R} \sum_{s \in \mathcal{S}} P_{\ell_i}(s \mid r) P_{\ell_j}(r \mid s) \tag{3.7}$$

where R stands for the full set of potential referents and $\mathcal{S}$ for the full set of signals. Using ca as a measure of informativeness allows us to categorise lexicons according to how much information they provide about the speaker's intended referent. Rounding informativeness scores down to two decimals leads to a classification of the 343 logically possible 3x3 lexicons into 14 categories (where category sizes range between 6 and 63 lexicons; with mean size 24.5). These categories will henceforth be referred to as *lexicon types*. Example lexicons from three of these 14 lexicon types are shown in Figure 3.3.



(a) Example lexicon from uninformative lexicon type (ca = 0.33...)

(b) Example lexicon from lexicon type with relatively low informativeness (ca = 0.51)

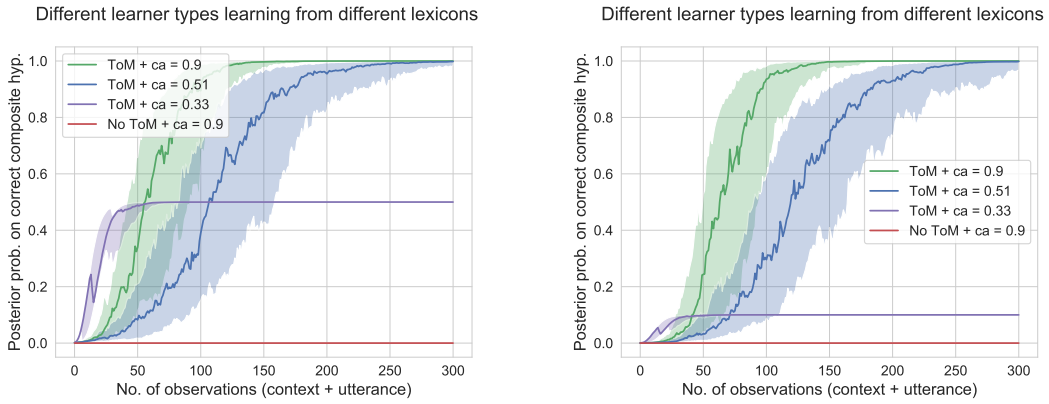
(c) Example lexicon from lexicon type with maximum informativeness (ca = 0.90)

Figure 3.3: Example lexicons from three of the 14 different lexicon types. In these matrices, referents are represented by rows and signals by columns. Blue squares represent an association between the corresponding referent and signal, while white squares represent the absence of such an association.

Each of the figures below shows learning results obtained when the learner receives input from a speaker who has the *opposite* perspective to the learner. Inferring the speaker's perspective is thus particularly challenging for the egocentric learner, who has to overcome an initial bias in favour of the assumption that the speaker will share their perspective. Each individual observation (i.e. data point d) that the learner is exposed to always consists of a randomly generated context combined with one utterance, where the utterance is generated based on the speaker's perspective and lexicon as described in Section 3.4.

Firstly, Figure 3.4 shows the learning results obtained with a selection of different combinations of learners and speakers. In order to investigate how learning is affected by the learner's perspective-taking abilities and the speaker's lexicon, figures 3.4a and 3.4b each compare four different conditions. In each condition, the learner receives input from a speaker who has the opposite perspective to the learner. What differs between conditions is either the informativeness of the speaker's lexicon, or the learner's ability to represent the speaker's perspective. The different speaker lexicons for which learning results are shown in Figure 3.4 are (i) a maximally informative lexicon, (ii) a lexicon with relatively low informativeness, and (iii) a completely uninformative lexicon (i.e. the three lexicons shown in Figure 3.3). The learner in these first three conditions is always a 'regular' learner, either unbiased or egocentric. In contrast, the fourth condition shows a learner who cannot represent the possibility that the speaker's perspective may be different from their own (i.e. who starts out with a prior probability of 0.0 assigned to the correct perspective hypothesis). This final learner receives input

from a speaker with a maximally informative lexicon.



(a) Unbiased learner

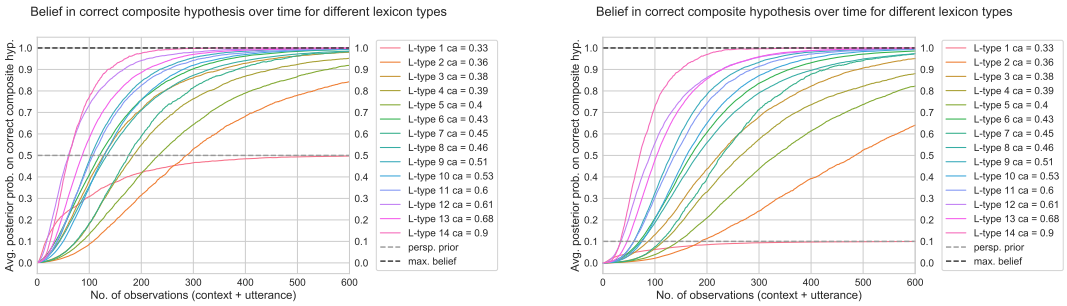
(b) Egocentric learner

Figure 3.4: Learning curves for different learner types receiving data from speakers with different lexicons. Subfigure a shows results for the unbiased learner, and subfigure b shows results for the egocentric learner. Each of these subfigures in turn shows learning results for (i) a learner receiving input from a maximally informative lexicon (in green), (ii) a learner receiving input from a low-informativeness lexicon (in blue), (iii) a learner receiving input from an uninformative lexicon (in purple), or (iv) a learner who cannot represent the possibility that the speaker's perspective may be different from their own, receiving input from a maximally informative lexicon (in red). *ca* stands for the communicative accuracy of the lexicon with itself, which is used as a measure of the lexicon's informativeness and can range between 0.33... and 0.90 (given lexicon size = 3x3 and error rate $\epsilon = 0.05$). *ToM* stands for theory of mind: a *ToM* learner has an egocentric bias but is able to represent the possibility that the speaker's perspective might be different from their own, while the *No ToM* learner cannot represent this possibility at all (i.e. for the *No ToM* learner the correct perspective hypothesis has a prior probability of 0.0). Learning curves show the amount of posterior probability assigned to the correct composite hypothesis (i.e. lexicon + perspective) over time (i.e. number of observed contexts). Lines show median and shaded areas show upper and lower quartiles over 100 independent simulation runs per condition.

Figure 3.4 shows that both the unbiased and the egocentric learner are able to fully learn the correct composite hypothesis if (i) they are able to represent the possibility that the speaker has a different perspective from their own, and (ii) the speaker's lexicon is not uninformative. Although the learner takes longer to fully acquire the correct composite hypothesis when the speaker uses an ambiguous lexicon compared to a maximally informative lexicon, the ambiguous input does not prevent full acquisition from happening. In contrast, when the speaker uses a completely uninformative lexicon, neither learner's belief in the correct composite hypothesis exceeds the prior probability they assigned to the correct perspective hypothesis, no matter how many contexts they observe (this result is analysed in more detail below). Finally, when the learners are not able to represent the speaker's true perspective (because they start out with 0.0 prior

probability on that hypothesis), neither learner accumulates any belief in the correct composite hypothesis. (This result is not surprising, given that no matter how likely the data is under this hypothesis, it will be multiplied with a prior probability of 0.0 to yield the learner's posterior belief in the correct hypothesis.)

To get a more complete image of how the speaker's lexicon affects the timecourse of learning, Figure 3.5 presents summary graphs of how the learner's belief in the correct composite hypothesis develops over time given each of the 343 possible input lexicons, averaged over lexicon types.



(a) Unbiased learner; composite hypothesis (b) Egocentric learner; composite hypothesis

Figure 3.5: Learning curves for learners with different perspective priors receiving data from a speaker that has the opposite perspective from the learner, for all different possible input lexicons, categorised by lexicon type. *ca* levels of different lexicon types range from lowest possible (0.33...) to highest possible (0.90) for lexicon size = 3x3 and error rate $\epsilon = 0.05$. Learning curves show amount of posterior probability assigned to correct composite hypothesis (i.e. lexicon + perspective) over time (i.e. number of observed contexts), averaged firstly over 100 independent simulation runs per input lexicon, and subsequently over all lexicons within a given lexicon type. Grey dashed line indicates the prior probability assigned to the correct perspective hypothesis. Black dashed line indicates maximum posterior probability that can be reached.

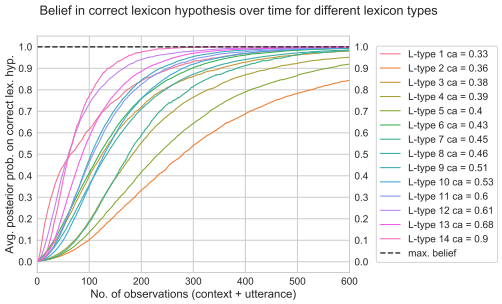
Firstly, Figure 3.5 shows that both types of learner can acquire each of the composite hypotheses, *except* when the speaker who provides input uses a fully ambiguous lexicon (in line with the results shown in Figure 3.4). This is lexicon type 1 with a *ca* score of 0.33..., which corresponds to chance level in a world with three referents, indicating that this lexicon type does not provide any information about the speaker's referential intentions. This lexicon type consists of lexicons which associate either one signal with all referents, two signals with all referents (where both signals individually map to each of the referents), or all signals with all referents. In Figure 3.5 we see that for this lexicon type the posterior probability assigned to the correct composite hypothesis never exceeds the prior probability that is assigned to the correct perspective hypothesis. This result is analysed in more detail below, where learning of the correct lexicon hypothesis

and learning of the correct perspective hypothesis are shown separately.

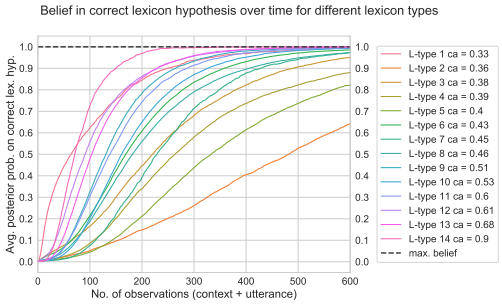
Secondly, comparing Figure 3.5a to Figure 3.5b shows that although the egocentric learner takes longer to acquire each of the correct composite hypotheses than does the unbiased learner, there is no striking qualitative difference in their learning results. Specifically, the ranking between the different lexicon types in terms of learning speed is roughly similar for the two learner types. Thirdly however, looking at the shape of the curves reveals that the informativeness of the speaker's lexicon has a bigger impact on learning for the egocentric learner than it does for the unbiased learner. Specifically, we can see that the learning curves of the egocentric learner span a wider space of learning rates (i.e. the lines 'fan out' more). Thus, we might conclude that the informativeness of the lexicon the speaker uses makes a bigger difference for egocentric learners.

The results shown in Figure 3.5 only show how the learner's belief in the correct *composite* hypothesis develops over time. As mentioned above, it is also useful to look at how the two component parts of this hypothesis are acquired, especially for illuminating what happens when the speaker uses an uninformative lexicon (lexicon type 1). Figure 3.6 therefore teases apart the posterior probability assigned to the correct *lexicon* hypothesis (collapsing on the two perspective hypotheses), and the posterior probability assigned to the correct *perspective* hypothesis (collapsing on all lexicon hypotheses).

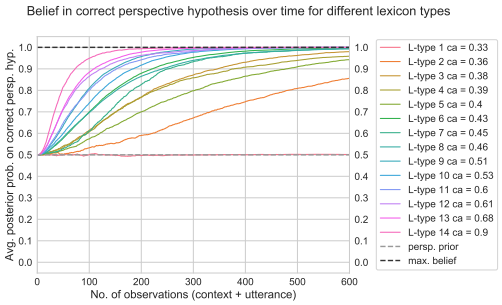
Firstly, if we look at lexicon-learning in isolation (subfigures 3.6a and 3.6b), we see that although the shape and ranking of the curves is mostly similar to those for the composite hypotheses shown in Figure 3.5, there is one striking difference: the learning curve for the uninformative lexicon type (type 1) reaches ceiling relatively quickly; second only to three or four of the most informative lexicon types. This reflects the fact that the uninformative lexicons are relatively easy to learn. As mentioned above, the class of uninformative lexicons (lexicon type 1) consists of those lexicons which associate either one signal with all referents, two signals with all referents, or all signals with all referents. Although the more signals the lexicon uses, the more observations it will take to learn, each of these subtypes can be learned without knowing anything about the speaker's perspective. Learning these lexicons only requires the learner to observe either that the speaker utters only one signal regardless of the context, or that the speaker utters two or three of the possible signals with equal frequency regardless of the context. That no information about the speaker's perspective needs to be gained in order to correctly learn these uninformative lexicons is evidenced by the fact that



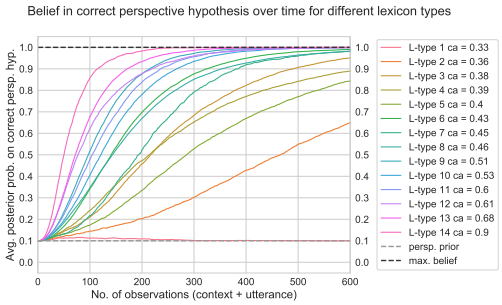
(a) Unbiased learner; lexicon hypothesis



(b) Egocentric learner; lexicon hypothesis



(c) Unbiased learner; perspective hypothesis



(d) Egocentric learner; perspective hypothesis

Figure 3.6: Learning curves for simulations shown in Figure 3.5, split by amount of posterior probability assigned to correct *lexicon* hypothesis (subfigures a and b) and amount of posterior probability assigned to correct *perspective* hypothesis (subfigures c and d). As in Figure 3.5, lines show grand means averaged first over 100 independent simulation runs per input lexicon, and subsequently over all lexicons within a given lexicon type. Grey dashed line indicates the prior probability assigned to the correct perspective hypothesis. Black dashed line indicates maximum posterior probability that can be reached.

the posterior probability assigned to the correct *perspective* hypothesis for this lexicon type, as shown in subfigures 3.6c and 3.6d, remains at the prior probability assigned to this hypothesis.

When we look at perspective-learning in isolation (subfigures 3.6c and 3.6d), we see firstly that the learners start out with a higher prior probability assigned to the correct hypothesis than when we look at learning of the composite hypothesis. This is due to the fact that there are only two possible perspective hypotheses, versus a total of 646 possible composite hypotheses (2×343), and is thus an artifact of how the learner's space of perspective hypotheses and lexicon hypotheses were chosen. Leaving the point of origin aside, we can see that the shape and ranking of the learning curves follow roughly the same pattern for learning about perspectives alone as they do for learning about composite hypotheses (as shown in Figure 3.5).

3.6 Discussion

In this chapter I have reviewed empirical evidence for the hypothesis that language and mindreading co-develop. As discussed in Section 3.1, controlled experiments with typically developing infants and children have shown that they readily use social cues to guide them in word learning, and suggest that they rely on an understanding of referential intentions to do so (see also Baldwin and Moses, 2001, for a review). The language development of children with autism spectrum disorder (ASD) seems to corroborate these findings, in the sense that ASD is associated with difficulties in understanding others' mental states, and that children with ASD show delayed and atypical language development (see also Eigsti et al., 2011; Tager-Flusberg et al., 2005, for further reviews). Word learning experiments similar to the ones used with typically developing children also indicate that children with ASD have difficulty learning word meanings if it requires inferring the speaker's referential intention, and may therefore have to rely on other strategies for learning language, such as contextual cues (Parish-Morris et al., 2007).

The other way around, as reviewed in Section 3.2, evidence from correlational, longitudinal and training studies indicates that language (and specifically being exposed to conversation about mental states) plays an important role in the development of mindreading abilities as well (see also Slaughter and Peterson, 2011, for a review). These findings are corroborated by the mindreading development of deaf children who grow up in hearing families, and therefore only start learning a sign language (or spoken language with the aid of hearing aids or a cochlear implant) by the time they enter primary school. These children show a delayed development of mindreading compared to both hearing children and their deaf peers who grow up with deaf parents and therefore start learning sign language from birth (see also Meristo et al., 2011; Peterson and Siegal, 2000; Pyers and de Villiers, 2013, for further reviews). Studies of the first cohort of signers of the emerging Nicaraguan Sign Language (NSL) and homesigners in Nicaragua who never learned a conventional sign language, show that having access to language and a 'community of minds' plays a crucial role in the development of full-blown mindreading.

In Section 3.3 I reviewed the extent to which the role of social cues and intention-reading have been incorporated in computational models of word learning for the purpose of reducing referential uncertainty. Two such models have shown good perfor-

mance on learning a lexicon of nouns from transcripts of videos of naturalistic mother-child interactions from the CHILDES corpus (MacWhinney, 2000): one by incorporating the social cues of prosody and joint attention (Yu and Ballard, 2007), and one by incorporating an ability to consider referential intentions as a mediating factor between the context and a speaker's utterances (Frank et al., 2009b). However, neither of these models takes into account the fact that the ability to infer referential intentions may itself need to be learned or developed, and that this development may depend in part on learning language.

Therefore, I presented a new model of word learning in Section 3.4 which implements vocabulary learning and learning about a speaker's perspective as a joint-inference task (see also Woensdregt et al., 2016, reproduced in appendix F). In this model, a speaker's linguistic behaviour is a result of an interaction between the speaker's individual perspective on the world, the current context, and the speaker's lexicon. However, the speaker's perspective and lexicon are not directly observable to the learner, who can only observe the speaker's utterances along with the contexts in which they occur. Thus, in order to learn words, the learner in this model needs to simultaneously infer both the speaker's perspective and the speaker's lexicon; the utterances that the learner observes are a result of the interaction between these two variables.

The learner in this model uses Bayesian inference in order to update their beliefs about which combination of perspective and lexicon hypothesis best explains the data observed. In Section 3.4 the design of the model is presented using the spatial metaphor that objects which are closer to the speaker in a given context will be more salient and therefore more likely to be chosen as the speaker's intended referent. However, the same design can equally model a more abstract situation where the speaker's perspective represents a 'view of the world' which, in interaction with a conversational context, influences how likely the speaker is to choose different topics of conversation.

The simulation results presented in Section 3.5 show that a Bayesian learner is able to solve the problem of having to jointly infer the speaker's lexicon and perspective as long as two conditions are met: (i) the learner is able to represent the speaker's perspective; and (ii) the speaker uses a lexicon that is not completely ambiguous. If the learner has a strong *un*helpful bias about the speaker's perspective (as in the case of an egocentric learner receiving input from a speaker who has the opposite perspective to the learner), this slows learning down somewhat, but does not prevent the learner from ultimately correctly inferring the speaker's lexicon and perspective. Similarly, the

more ambiguous the lexicon that the speaker uses, the longer the learner will take to infer the speaker's perspective and lexicon. However, as long as there is at least *some* information in the lexicon (just one signal that doesn't refer to all objects is enough), the learner will ultimately get there.

Learning about the lexicon and learning about the speaker's perspective thus go hand-in-hand. This is not surprising given the design of the model: because the speaker's utterances are a result of an interaction between the speaker's perspective, the current context, and the speaker's lexicon, a learner who is able to take the speaker's perspective gains information about which object the speaker's utterance is likely to refer to (assuming the learner can observe the context). And vice versa, if the learner knows the lexicon, this provides information about the speaker's referential intentions in different contexts, and therefore about the speaker's perspective.

Somewhat more surprising perhaps is the finding that these skills can be bootstrapped from each other from scratch. As mentioned briefly in Section 3.4, this is a result of the fact that because the model treats the saliencies of referents as relative, two opposite perspectives on the same context will nearly always lead to two different probability distributions over potential referents, even when the 'order' of referents is ignored (i.e. the ratios between the probabilities within each distribution will be different). Thus, a Bayesian learner has a 'way in' to gaining information about the speaker's perspective that is not dependent on knowing which signal maps to which referent, simply by observing the frequencies with which the speaker uses different signals in different contexts. However, this baseline strategy works best if the speaker uses a different and unique signal for each possible referent. As the speaker's lexicon becomes more ambiguous, the learner will need to observe more data in order to gain information about the speaker's perspective. This explains the positive correlation between the 'informativeness' of the speaker's lexicon and the rate of learning reported above. It also explains why a learner who receives input from a speaker with a completely uninformative lexicon will never gain more information about the speaker's perspective than what was captured in their a priori assumptions, as shown in Section 3.5.

The case of a learner receiving input from a speaker who uses a very ambiguous lexicon can be likened to the situation of a learner who is trying to divine something about other minds simply through observing others' non-linguistic behaviour. Although this strategy should get an observant learner some way towards understanding that others' mental states can be different from her own, the data she would rely on for

learning this would be far sparser and more ambiguous than the evidence received by a learner born into a community that uses a fully-fledged conventional language. The simulation results presented in this chapter suggest that the more the lexicon that the learner receives input from is optimised for communication, the more information it provides about the speaker's perspective, which in turn increases the ease with which the learner can infer the speaker's perspective. This dynamic results from the simple assumption that a speaker's utterances are not a direct result of the context, but rather of an interaction between the context and the speaker's perspective on the world, which is 'hidden', subjective variable. Under this assumption, a learner who receives input from a lexicon that does not provide any information about the speaker's referential intentions will never learn what the speaker's perspective is (unless this is known a priori).

These simulation results relate to the different hypotheses that have been put forward regarding *how* language promotes the development of mindreading, as discussed in Section 3.2.1. To recap, these hypotheses are that (i) language provides labels for mental states, which help learning and reasoning about them (Olson, 1988); (ii) language provides specific grammatical structures (sentential complement constructions) which provide a conceptual framework that helps learners think about mental states (de Villiers and Pyers, 2002); and (iii) language provides general exposure to evidence of diverging perspectives and draws learners' attention to others' mental states (Harris, 1996). Note that these three hypotheses are not mutually exclusive; several or even all of these mechanisms may contribute to the development of mindreading together.

Some initial empirical evidence has been found in favour of each of these three hypotheses. As discussed in Section, 3.2.2, a longitudinal study by Pyers and Senghas (2009), following the first two cohorts of NSL signers, showed that improvements in signers' false belief understanding over time were always preceded by, or otherwise co-occurred with, an increase in their use of mental state verbs, providing evidence in favour of the first hypothesis. As discussed in Section 3.2.1, a training study by Lohmann and Tomasello (2003) showed that both children in a *sentential complement only* condition and children in a *discourse only* condition improved in their false belief understanding, relative to children in a *no language* condition. These findings provide evidence in favour of the second and third hypothesis respectively. Furthermore, Lohmann and Tomasello found that a *full training* condition which combined sentential complements and discourse yielded the biggest improvements in children's false belief

understanding, suggesting that the effects of these two mechanisms are complementary.

The modelling work presented in this chapter adds to this initial empirical evidence in favour of the third hypothesis, in the sense that it provides a proof of concept demonstration that discourse itself can provide learners with opportunities to learn about differing perspectives. That is, learners in the model presented here do not specifically learn mental state terms or sentential complement constructions, but learning a lexicon nevertheless helps them infer a speaker's perspective on the world, and vice versa.

One empirical finding that seems to challenge the hypothesis that linguistic input is necessary for the development of full-blown mindreading, is the evidence of implicit false belief understanding found in infants (Onishi and Baillargeon, 2005; Southgate et al., 2007). As discussed in Section 3.2, these findings are currently under debate due to several failures to replicate (see Schuwerk et al., 2018; Powell et al., 2018, and the rest of the special issue of *Cognitive Development* titled "Understanding theory of mind in infancy and toddlerhood", vol. 46, 2018). However, even if future work shows that the finding of implicit false belief understanding in infants does hold true[6], this would not necessarily contradict the hypothesis that language is important for the development of mindreading.

The discrepancy between implicit mindreading potentially coming online early and explicit mindreading coming online much later can be elegantly accounted for by a dual process model. Two examples of such a model of mindreading were discussed in Section 3.2: the *two-systems* account (Apperly and Butterfill, 2009; Butterfill and Apperly, 2013), and the *submentalizing* account (Heyes, 2014b,a, 2015, 2018). Both these accounts state that whereas implicit mindreading abilities result from genetically inherited cognitive processes, the development of explicit mindreading abilities is likely to depend on language (Apperly and Butterfill, 2009; Heyes, 2014b; Heyes and Frith, 2014). Thus, according to these accounts, the finding of implicit mindreading abilities in infants is not incompatible with the hypothesis that exposure to language is necessary for developing full-blown, explicit mindreading abilities.

As summarised above, the simulation results presented in this chapter showed that co-development between word learning and perspective-learning can only get off the ground if the speaker uses at least a somewhat informative lexicon. This leads to the

[6] One initiative that is expected to shed further light on this question is ManyBabies project 2 (Frank et al., 2018).

question how a *population* of such agents could establish an informative lexicon from scratch (i.e. if they start out with a completely uninformative one). This question will be addressed in the next chapter, by incorporating the developmental model described here in an iterated learning model.

The model presented in this chapter has several limitations in terms of how it addresses the empirical literature reviewed in sections 3.1 and 3.2. Firstly, it is important to note that the *ability* to learn about perspectives is not learned in this model, but is innately specified. That is, learners start with an innate knowledge of how a given perspective combined with a given context would give rise to a given saliency distribution over referents. This function does not need to be learned. The only thing that the learner has to infer based on the data is which of the two possible perspective hypotheses applies to the speaker, where the speaker's perspective is a single unknown parameter in the innately given function that maps from a publicly available context to a speaker-specific saliency distribution over potential referents.

Secondly, there is an asymmetry in the model between speakers and listeners: learner agents are able to represent that the speaker has a subjective perspective on the world which influences their utterance production. However, speaker agents (who in an iterated learning version of this model would have once started out as learners themselves) do not take into account that the listener/learner does such perspective-taking when choosing which signal to produce; they simply produce signals literally based on their own lexicon. This issue is addressed in Chapter 5, in which pragmatic reasoning is added to the model of communication.

Chapter 4

Cultural evolution of lexicons in populations of perspective-taking agents

Chapter 3 presented a new Bayesian model of co-development lexicon- and perspective-learning. In this model, the learner is faced with the joint inference task of having to simultaneously learn about a speaker's lexicon and perspective, without receiving direct evidence of either. The only data the learner gets to observe consists of the speaker's utterances in context. From this the learner has to infer the speaker's perspective and lexicon, which are unobservable but both independently influence what utterances the speaker produces in different contexts. The simulation results obtained with this model as described in Chapter 3 showed that a Bayesian learner can solve this joint inference problem by bootstrapping one skill (lexicon-learning) from the other (learning about the speaker's perspective) and vice versa, even if the learner starts out with an unhelpful egocentric bias. However, this process of co-development was only successful if the learner received input from a speaker whose lexicon is at least somewhat informative (i.e. not completely ambiguous).

This thesis is ultimately concerned with language evolution rather than development, and specifically whether language emergence in the hominin lineage may have gone hand-in-hand with the evolution of more sophisticated perspective-taking skills. The co-developmental results summarised above lead to the question how a population of agents that develop in this way could evolve informative lexicons from scratch. That is, if individual learners require input from an informative lexicon in order to correctly

infer a speaker's perspective, and correctly inferring the speaker's perspective is necessary for the faithful transmission of the lexicon, how could a population of such agents establish an informative lexicon if they start out without any meaningful signal-referent mappings?

In this chapter I will address this question by incorporating the developmental model described in Chapter 3 into a model of *iterated learning*, in which agents learn their lexicon by observing the utterances of agents from a previous generation who have learned their lexicon in the same way. I will explore the extent to which different selection pressures facilitate the emergence of informative lexicons in the population, and how this in turn affects agents' success at communicating with each other and inferring each other's perspectives. This chapter investigates to what extent a co-evolution between the emergence of informative lexicons and agents' perspective-taking abilities can be the result of cultural evolution alone; that is, without agents' innate ability to learn either skill changing over generations. I will first review the findings of existing models of cultural evolution in Section 4.1, followed by a description of the current model in Section 4.2. Section 4.3 presents the simulation results obtained with this model of iterated learning of lexicons and perspectives, exploring how different selection pressures influence what types of lexicons evolve in the population. Finally, Section 4.4 reviews the findings of this model in relation to the findings of other work on cultural and language evolution, and in relation to the aims of this thesis.

4.1 Review of existing models of cultural evolution

4.1.1 The iterated learning model

The *iterated learning model* describes the process by which a behaviour is culturally transmitted over generations through social learning (Kirby, 2001). This process qualifies as iterated learning if the behaviour is acquired through a process of induction based on observations of that behaviour in another individual who has acquired the behaviour in the same way (Kirby et al., 2014). Thus, the behaviour is passed along a transmission chain of individuals, and each new learner (or group of learners in the case of a population) can be thought of as a new generation. Individuals of a new generation arrive at their own internal model of the behaviour by observing the externalised behaviour of one or more individuals from the previous generation, and subsequently externalise the behaviour, which provides data for learners in the next generation (see

Figure 4.1). The iterated learning model has been widely used to simulate language evolution both in computational models and laboratory experiments (for reviews see Kirby et al., 2014, 2015; Kirby, 2017; Smith, 2018).

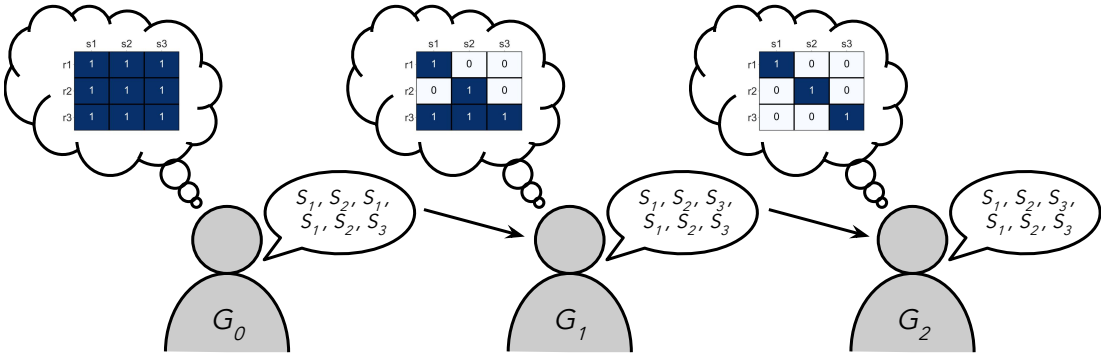


Figure 4.1: Diagram of iterated learning using the model of lexicons and utterance production presented in Chapter 3 as an example. An agent from generation G_t produces data based on their own lexicon; this data is then observed by a learner from the next generation (G_{t+1}), who induces their own lexicon based on the data. After lexicon induction, the agent from generation G_{t+1} produces data for generation G_{t+2}, and so on.

Because every individual in the iterated learning model acquires the behaviour through induction on observations, the behaviour is not simply copied from one generation to the next, but reconstructed by each new individual. In the case of language, we know that real-world language learners have to reconstruct the language of their speech community based on a finite subset of data. In the iterated learning model this is instantiated by what is known as the *bottleneck* on cultural transmission: each learner observes only a subset of all data that could be produced with the language, and for a language to be transmitted faithfully it needs to pass through this transmission bottleneck (Kirby, 2001). Thus, languages that are more easily reconstructable from a limited set of data have a higher chance of persisting over generations. Agent-based models of iterated language learning have shown that this transmission bottleneck can cause an initially unstructured (e.g. holistic) language to become structured (e.g. compositional) over generations, because such systematic structure as compositionality enables learners to correctly generalise a full linguistic system from a finite set of data (Brighton, 2002; Brighton et al., 2005; Kirby, 2000, 2001, 2002; Zuidema, 2003).

Aside from the transmission bottleneck, another factor that influences which languages endure the process of iterated learning is the inductive bias of the learners. This inductive bias determines the ease with which learners acquire the different languages.

Computational models of iterated learning with inductive biases have demonstrated that such biases can lead to small transformations of the behaviour being transmitted, which can accumulate over generations such that the behaviour is ultimately drastically changed to fit the learners' bias (see Kirby et al., 2014, 2015; Kirby, 2017; Smith, 2018, for reviews). For instance, Kirby et al. (2007) showed with a Bayesian version of the iterated learning model that if agents have a learning bias in favour of regularity, transmission chains will predominantly converge on regular languages, no matter what type of language the first generation started out with.[1] Moreover, Kirby et al. (2007) demonstrated that the strength of the learners' regularity bias does not affect the extent to which regular languages are overrepresented: cultural transmission amplifies the biases of individual learners, no matter how weak, to ultimately create a strong effect on the distribution over languages.

In contrast with the strength of the bias, Kirby et al. (2007) showed that the width of the transmission bottleneck does influence how likely regular languages are to become dominant. Specifically, the fewer observations per learner (i.e. the tighter the bottleneck), the stronger the effect of the learners' bias. This result makes intuitive sense if we consider how bias and data interact in the Bayesian inference model of learning. As described in Chapter 3 (Section 3.4), Bayesian inference combines two quantities: the prior probability of a given hypothesis (which represents the learner's belief in that hypothesis prior to seeing any data; the learner's bias), and the likelihood of the observed data under the hypothesis. The more data the learner observes, the more the likelihood of the data given the correct hypothesis will increase relative to that given other hypotheses, and thus the stronger the effect of the data on the learner's posterior probability distribution. Therefore, the *less* data the learner sees, the more influence the prior bias will have, because the less its effect will be 'washed out' by the data.

Agent-based models of iterated learning can be implemented using any model of individual learning (see Kirby et al., 2014; Kirby, 2017; Smith, 2014, for reviews). However, as demonstrated by Griffiths and Kalish (2007), using Bayesian inference as a learning model comes with the advantage of making the inductive bias of the learners fully explicit. This is important if we want to assess exactly how the process of

[1]This result holds only if learners do not select a hypothesis (i.e. language) with a probability exactly proportional to its posterior probability (known as *sampling*), but instead disproportionately favour languages with higher posterior probability (known as *maximum a posteriori* or MAP selection). I discuss the effect of these different hypothesis selection methods in more detail below.

cultural transmission (as opposed to biological adaptation) can influence the structure of language, because this requires us to 'partial out' the contribution of individual learners, in order to reveal the cumulative effect of cultural transmission (Kirby, 2017). In the Bayesian inference model of learning, learners' inductive bias, and thus the individual learners' contribution to the process of iterated learning, is captured fully and explicitly by their prior probability distribution over hypotheses.

Griffiths and Kalish (2007) demonstrated that in such Bayesian models of iterated learning, the added effect of cultural transmission depends on the way in which learners select a hypothesis (language in this case) after learning. Griffiths and Kalish explored two methods of hypothesis selection: *sampling*, where learners choose a hypothesis with probability exactly equal to its posterior probability, and *maximum a posteriori* (MAP) selection, where learners choose the hypothesis with the highest posterior probability. Griffiths and Kalish showed that when learners use sampling, the result of iterated learning is determined entirely by their prior bias. That is, the distribution over languages resulting from learners' hypothesis selection will eventually come to be exactly equal to the learners' prior probability distribution over languages. Thus, when learners sample, iterated learning will simply reveal the learners' inductive bias, and the process of cultural transmission does not add anything on top of that.[2]

When learners use the MAP hypothesis selection strategy in contrast, cultural transmission amplifies their bias. That is, the ultimate distribution over languages will be an exaggerated version of the learners' prior bias, as shown by Kirby et al. (2007) for the case of a bias in favour of regularity. In fact, Kirby et al. (2007) parameterised learners' hypothesis selection method with a parameter interpolating between pure sampling and pure MAP selection, and found that in all cases where it is skewed towards MAP selection, regular languages end up being systematically overrepresented. Finally (and again in line with the results of Kirby et al. (2007)), Griffiths and Kalish showed that in the case of MAP selection, the effect of learners' inductive bias gets amplified more strongly if there is a tighter bottleneck on transmission. In the case of sampling, the width of the bottleneck does not affect the ultimate distribution over languages (although it does affect how long transmission chains take to reach that stationary distribution; with

[2]Some exceptions to this convergence to the prior result have been demonstrated, such as when populations are heterogeneous in terms of agents' bias strength (Navarro et al., 2018), or when learners receive input from multiple cultural parents (Smith, 2009) (although Burkett and Griffiths, 2010 show that the latter only causes non-convergence to the prior under the non-rational assumption that learners try to infer a single language while in fact receiving input from speakers of different languages).

tighter bottlenecks leading to quicker convergence).

4.1.2 The interplay between learning biases and selection

The effect of learners' inductive biases on the cultural evolution of languages as described above works by means of *convergent transformation* (Claidière et al., 2014a). That is, each learner reconstructs the language from a limited amount of input, and the learner's inductive bias causes the language to be transformed in a direction that makes it more learnable. If learners are biased, these transformations are not random but directed. And when all learners in a population share the same bias, these transformations will be convergent as well, causing languages to systematically move in the direction that is favoured by the bias. This holds not just for the evolution of language, but for the evolution of any cultural trait that involves biased reconstruction during transmission. This process of convergent transformation is also known as cultural *attraction* (Sperber, 1996; Claidière et al., 2014a; Scott-Phillips et al., 2018), where the central idea is that because different possible transformations are not equally probable, they will be biased in a certain direction, and this direction acts as an *attractor* on cultural evolution.

A different mechanism that can lead to changes in cultural traits over generations is cultural *selection*. This consists of individuals selecting which variant of the cultural trait to copy, or who to copy from. In their seminal work on adapting population genetics models to the case of cultural evolution, Boyd and Richerson (1985) distinguished three main types of such cultural selection: trait-based (originally named 'direct bias' and later 'content bias'), model-based (originally named 'indirect bias') and frequency-based (originally named 'frequency-dependent bias'). In trait-based cultural selection, individuals select which variant of a trait to copy based on characteristics of the variant itself, such as its effectiveness or efficiency. In model-based cultural selection, individuals instead select who to copy from (i.e. who they want to be their model), based on characteristics of the model individual, such as success or prestige.[3] Finally, in frequency-based selection, individuals select which variant to copy based on its frequency in the population. An example of this type of selection is conformism, in which individuals preferentially copy the majority variant.

[3]Model-based selection can serve as a heuristic for selecting on the success of the variant itself (when the latter is hard or time-consuming to assess), assuming that the model's success or prestige stems at least in part from the variant she uses.

In contrast to convergent transformation, selection can change the frequency of cultural variants over generations also when transmission is replicative rather than reconstructive. In fact, selection can only be effective under low levels of transformation (i.e. low levels of 'mutation', to use the biological selection equivalent), because the process by which individuals select a variant can only have a population-level effect if the variant that the individual ends up with retains the characteristics that made it successful in selection. In other words, cultural selection relies on transmission being high-fidelity. Also in contrast to convergent transformation, selection relies on variation being present in the population, because it works by means of individuals choosing among existing variants, rather than among 'internally-generated' variants as in the case of convergent transformation (Boyd and Richerson, 1985, chapter 5).

Both convergent transformation and selection can lead to cultural stability, although the underlying mechanism is different. Under convergent transformation, stability arises when a cultural variant has emerged that sits close to an attractor (i.e. fits well with the biases of the individuals), after which subsequent transformations are unlikely to move away from this variant. Under selection in contrast, cultural stability is achieved when the probability of transformations is low and selection can therefore increase the frequency of the successful variant(s) at the cost of the less successful ones.

Importantly, convergent transformation can generate variation, while selection cannot. That is, convergent transformation generates new variants which will tend in the direction of the attractor, while selection can only choose between existing variants and is dependent on infrequent transformations for new variants to be generated. (As mentioned above, these mutations have to be infrequent because selection can only be effective if transmission is relatively high-fidelity.) The population-level consequence of this difference is that the strength of the effect of selection depends on the amount of variation present in the population, while the effect of convergent transformation does not (Boyd and Richerson, 1985, chapter 5). The relative importance of the processes of convergent transformation (i.e. attraction) and selection in cultural evolution has been a matter of much recent debate (see e.g. Acerbi and Mesoudi, 2015; Henrich et al., 2008; Scott-Phillips et al., 2018), as well as the question how the two processes interact. The same cultural trait can be subject to both convergent transformation and selection at the same time, and the direction of these two forces does not necessarily coincide. Below I will review the results of models and experiments developed to explore what happens in such cases.

Henrich and Boyd (2002) showed with a mathematical model (Model 1 in the paper) that as long as a cultural trait has multiple attractors and these attractors are relatively strong, the ultimate equilibrium of the system is determined by selection alone. They modelled a cultural trait x as a continuous one-dimensional space of variants, and looked at what happens when this trait is subject to two attractors at both extremes of the space ($x = 0$ and $x = 1$), as well as a weak selection pressure towards one of the extremes ($x = 1$). They further assumed that attraction is deterministic: there is a threshold value m in the space of x such that all variants $x < m$ are transformed in the direction of attractor $x = 0$, and all variants $x > m$ are transformed in the direction of attractor $x = 1$. In other words, each of the attractors has a discrete 'domain of influence', and whenever an individual chooses to learn from a model whose variant falls within the domain of a given attractor, the variant of the individual will necessarily transform in the direction of that attractor.

Henrich and Boyd (2002) demonstrated that when the two attractors are sufficiently strong relative to the selection pressure, the final equilibrium of the trait is determined entirely by selection. With such strong attractors, variants that start in the middle of the trait space are quickly transformed towards one of the attractors. Subsequently, once a variant is in the vicinity of an attractor, transmission becomes very faithful because transformations away from an attractor are much less likely than transformations towards it. Henrich and Boyd show that after this initial period of attraction-driven evolution, the longer-term dynamics and the final equilibrium of the system can be approximated with the standard discrete-trait replicator model. This model is used to describe the evolution of traits under selection alone, when attraction does not play a role. The stronger the attractors, the better the evolutionary dynamics of the system are approximated by this discrete-trait replicator model. In other words, given sufficiently strong attractors the process of attraction essentially 'self-eliminates' within an initial time period, after which the variants corresponding to the two attractors turn into replicators, and the subsequent evolution of the trait is determined entirely by selection. [4]

Claidière and Sperber (2007) point out that Henrich and Boyd (2002)'s assumptions that (i) the variant favoured by selection coincides with an attractor and (ii) attraction is deterministic, are not necessarily realistic. Regarding the first assump-

[4]Henrich and Boyd (2002) report that the same results hold for multi-dimensional traits and traits with more than two attractors.

tion, Claidière and Sperber (2007) argue that it in many cases of cultural evolution, learners tend to choose the more skilful individuals as models, even though they might end up with a simpler or otherwise less admirable version of the trait themselves; a case of selection and attraction working in opposite directions. Regarding the second assumption, Claidière and Sperber (2007) argue that it departs from Sperber's (1996) original definition of attraction in terms of a greater *probability* of transformations towards, rather than away from, a given point or area in the space of possible variants (i.e. the attractor), which views attraction as a probabilistic process.

Claidière and Sperber (2007) therefore extended the Henrich and Boyd (2002) model to explore what happens when these two assumptions are relaxed. To relax the first assumption (that the variant most favoured by selection coincides with one of the attractors), Claidière and Sperber modelled selection as a Gaussian distribution around a point in the trait space ($x = 0.7$), instead of a linear function as in Henrich and Boyd (2002). Simulation results obtained with this model show that the distribution over traits in the population converges to $x = 1.0$ rather than $x = 0.7$. That is, instead of converging on the variant favoured by selection, populations converge on the attractor nearest to the variant favoured by selection. Thus, when selection does not coincide with an attractor, *both* selection and attraction influence the ultimate equilibrium of the system, and the final distribution over variants shows the combined effect of these two forces. As Henrich et al. (2008) point out however, this means that the ultimate dynamics and equilibrium of the system is still well-described by replicator dynamics. That is, both attractors turn into replicators after an initial period of attraction-driven evolution, and the attractor-replicator which most increases individuals' success (i.e. which is closest to the peak of the Gaussian selection force) ultimately spreads through the population in a process that is captured by the discrete-trait replicator model.

To relax the second assumption (deterministic attraction), Claidière and Sperber added stochasticity to the model of attraction by letting transformations be chosen randomly from a range around the value that the variant would transform into in Henrich and Boyd's (2002) model. Thus, when the variant that a learner is trying to acquire is not too far from the edge of the domain of influence of one attractor, the learner might end up with a variant that is in the domain of influence of the other. This model gives rise to a distribution over variants in which all variants have a nonzero representation, and the shape of the distribution is determined by the relative strength of the two attractors. Finally, when such stochastic attraction is combined with a

Gaussian selection force peaking around $x = 0.7$, thus relaxing both assumptions in the same model, the ultimate distribution over variants is determined by both attraction and selection. When the strength of selection is increased, the dynamics of the system come closer to replicator dynamics, but equilibrium distribution over variants can still not be accounted for by selection alone.[5]

Griffiths et al. (2008) explore the interaction between selection and attraction in a Bayesian iterated learning model with two hypotheses. In this model, attraction is instantiated by learners' prior bias favouring one of the two hypotheses. Selection on the other hand is implemented as unequal fitness of the two hypotheses. That is, under selection for hypothesis 1, individuals who have adopted hypothesis 1 are more likely to provide input for the next generation than individuals who adopted hypothesis 2. Griffiths et al. show that when attraction and selection work in opposite directions, the resulting distribution over the two hypotheses is determined by both attraction and selection under a limited set of values of noise in transmission and bias strength. Under this limited set of parameter settings, selection pulls the distribution over hypotheses in the direction of the hypothesis that it favours. As selection strength increases, this effect increases, but never to such an extent that the hypothesis favoured by selection turns into a replicator and the equilibrium distribution is determined by selection alone. (Unless, presumably, the noise in transmission is reduced to 0.0, but this case is not specifically reported by Griffiths et al..)

However, Griffiths et al. (2008) go on to show that there is also a broad range of combinations of noise level and bias strength under which selection has no effect on the equilibrium distribution whatsoever. The stronger learners' prior bias in favour of one of the hypotheses, and the higher the level of noise in transmission, the less likely selection is to have an effect. (Where noise in transmission is defined as the probability that an agent with hypothesis i will produce a data point that 'belongs to' hypothesis j.) Thus, under the assumptions of a Bayesian iterated learning model where agents sample from their posterior probability distribution to select a hypothesis, the equilibrium distribution over hypotheses can be determined entirely by convergent transformation in a broad range of circumstances. This is in line with the results of Claidière and Sperber (2007) showing that when attraction is changed from being deterministic to

[5]See Henrich et al. (2008) for a response to the final results of Claidière and Sperber (2007), discussing mechanisms by which individuals can 'cut through' noise on transmission (such as stochastic attraction), with the effect of bringing cultural evolution closer to replicator dynamics again.

being probabilistic, the influence of attraction on the equilibrium distribution increases.

Taken together, these model results show that the interplay between convergent transformation and selection depends on a range of factors, including the level of noise in transmission and the relative strength and direction of convergent transformations and selection. There is both a range of cases under which the equilibrium distribution over variants is determined solely by selection (as demonstrated by Henrich and Boyd, 2002), and a range of cases under which the equilibrium distribution is determined solely by convergent transformation (as demonstrated by Griffiths et al., 2008). And finally, there is a range of cases in between in which the outcome of cultural evolution is a middle ground between the two forces (as demonstrated by Claidière and Sperber, 2007, and Griffiths et al., 2008).

Empirical evidence of this conclusion is provided by Claidière et al. (2018) in an iterated learning experiment with baboons (*Papio papio*). In this study, baboons were first trained to reproduce visual patterns of four squares lighting up in a grid of 16 squares, after which they took part in an iterated learning experiment in which the output patterns of one baboon formed the input for the next. In a first study, Claidière et al. (2018) showed that baboons' reproduction of the patterns was both very low-fidelity (with patterns having an average chance of 20% of being transmitted faithfully), and strongly biased in the direction of mathematically rare configurations known as 'tetromino': compact shapes of four connected squares. The frequency of such tetromino in the system increased over generations of the transmission chain, which went hand-in-hand with an increase in performance (see Claidière et al., 2014b for the full report of this initial study without selection). Thus, in the absence of any selection pressure, the patterns resulting from cultural evolution are determined by convergent transformations resulting from a bias in favour of tetromino that is shared by different baboons in the population.

In a second study, Claidière et al. (2018) explored how these convergent transformations would interact with a selection pressure. Specifically, Claidière et al. compared the effects of a selection pressure that is *aligned* with baboons' transformations (favouring more compact patterns) with a selection pressure *opposite* to baboons' transformations (favouring a maximal distance between the squares in the pattern). Claidière et al. simulated the outcome of cultural evolution with both convergent transformation and selection using a two-step procedure. First, they estimated the transition matrix (i.e. the probabilities of each possible pattern transforming into each other pattern) from

the experimental data, in order to simulate convergent transformation. Subsequently, they simulated transmission chains in which at each time step, after convergent transformations had taken place, only a subset of the resulting patterns was selected to be used as data for the next generation, based on the respective selection pressure. This simulation procedure allowed Claidière et al. to compare different combinations of the presence and absence of convergent transformation and selection.

Focusing specifically on the case where the directions of convergent transformation and selection are opposite to each other, Claidière et al. showed that the effect of selection is stronger when combined with convergent transformations compared to random transformations. This was the case even though convergent transformations work in the opposite direction (favouring compactness) to selection (favouring maximal inter-square distance), and the selection pressure was very strong (only the 10 'best' patterns out of a total of 50 output patterns were selected as input for the next generation). With the low level of transmission fidelity found in the empirical data, transformations are very likely to happen from one generation to the next. When those transformations are random, the effect of selection is mitigated because the property that made the selected variant successful (maximal inter-square distance, in this case) is unlikely to be retained after transformation, thereby preventing the accumulation of structure. In contrast, when transformations are directed and convergent (because the bias is shared among members of the population), the selected patterns will transform into patterns that remain in the vicinity of the selected ones, thereby allowing structure to accumulate.

The study of Claidière et al. (2018) thus illustrates empirically that both convergent transformation and selection can simultaneously contribute to the outcome of cultural evolution. Furthermore, it demonstrates that with the realistically low level of transmission fidelity resulting from empirical data, selection in fact has very little effect unless it is combined with convergent (as opposed to random) transformations. Claidière et al. further illustrate this finding by extending the model of Henrich and Boyd (2002) (described above) to compare the situation where transformations are convergent (as explored by Henrich and Boyd) to a situation where transformations are random. This comparison shows that selection has almost no effect when transformations are not directed.

Taken together, the models and empirical studies reviewed in this section illustrate that both convergent transformation and selection are important for the outcome of

cultural evolution, and that the relative importance of transformations depends on the fidelity of transmission (which in turn depends on both the noise in transmission and the strength of the learners' biases). The higher the probability of transformations (i.e. the lower the transmission fidelity), the more selection depends on those transformations being convergent (i.e. directed by a bias that is shared among all members of the population). Given the same level of fidelity, the effect of selection can be virtually cancelled out by random transformations, while being maintained when transformations are convergent; even if the direction of those convergent transformations is opposite to the direction of selection. Convergent transformations enhance the effect of selection by bringing variants closer to attractors, at which point the rate of transformation no longer disables selection because transformations remain close to their original. In such cases, it is the fact that transformations are convergent, rather than transformations being infrequent, which creates the cultural stability that is necessary for selection to operate.

4.1.3 Models of the role of joint attention in language evolution

To my knowledge, no other computational models of the co-evolution of language and perspective-taking (in the sense of inferring an internal, unobservable state of another agent) have been published to date. However, two models have been published which explore the role of joint attention in language evolution. Joint attention forms a precursor to perspective-taking and mindreading in typically developing children (Charman et al., 2000; Moore and Corkum, 1994), and as discussed in Chapter 3 (Section 3.1), the extent to which children engage in joint attention is a good predictor of language development in both typically developing and autistic children. Comparative research has also shown that where human children readily engage in and initiate joint attention with others, nonhuman primates do not (Tomasello et al., 2005; Tomasello and Carpenter, 2007). Tomasello et al. (2005) and Tomasello and Carpenter (2007) further argue that joint attention plays a crucial role in enabling *shared intentionality* (the ability to engage with others in a collaborative interaction with a shared goal), and that such shared intentionality is key to human culture, including the evolution of language.

Computational models of the role of joint attention in language evolution have focused on the problem of *referential uncertainty* in word learning as described in Chapter 3 (Section 3.3). In short, referential uncertainty refers to the fact that a novel

word could refer to many different things present in the context in which it is uttered. This forms a problem for word learning because a language-learner does not have direct access to the meaning the speaker has in mind when she utters the word, and instead has to infer this meaning from the context. Section 3.3 in Chapter 3 reviewed several heuristics that have been shown to help narrow down the space of possible meanings. One of the heuristics discussed in this respect is joint attention, which Yu and Ballard (2007) added on to an associative word learning model in order to 'highlight' relevant objects in the context. Yu and Ballard (2007) showed that an associative learning model with joint attention performed better at learning a lexicon from naturalistic data than without, and even better if combined with prosodic cues which highlight words in the speech-stream. Kwisthout et al. (2008) and Gong and Shuai (2012) also modelled joint attention as a mechanism for reducing referential uncertainty, but focused on its role in language emergence and evolution. Kwisthout et al. (2008) modelled three different forms of joint attention and explored how each may be relevant in language emergence, while Gong and Shuai (2012) modelled the potential co-evolution of joint attention and language. I will discuss these two models in turn below.

Kwisthout et al. (2008) model three different forms of joint attention which correspond directly to three different stages of joint attention development in children. These are, in developmental order, (i) *checking attention*, in which the child checks whether her caregiver is attending to the same object as her, (ii) *following attention*, in which the child allows her attention to be directed to an object by her caregiver; and (iii) *directing attention*, in which the child directs her caregiver's attention to an object (Carpenter et al., 1998). Kwisthout et al. add these mechanisms to a simple cross-situational word learning model in order to explore how each type of joint attention can help reduce referential uncertainty, and how this in turn facilitates language emergence in a group of agents.

Kwisthout et al.'s (2008) model is based on the *language game model* (e.g. Steels, 1996; Smith, 2014), which investigates how a group of agents can develop a shared language (i.e. converge on a shared set of form-meaning mappings). In Kwisthout et al.'s model, no new agents are added to the original group, so the model does not include generational turnover. In that sense, it is a model of language emergence more than language evolution. The way Kwisthout et al. represent languages is also inspired by the language game model, in which the meaning of a signal is not equated directly with an object in the context, but rather with a *feature* of an object. Thus, the same

meaning can be true of several different objects in the context if those objects share the corresponding feature. Specifically, each object consists of three attributes, which can each have four distinct values (objects can thus be thought of as a collection of features such as [LARGE, GREEN, SQUARE]). An agent's lexicon then simply consists of an association matrix between these features and a set of signals.

Each interaction between a speaker and a hearer happens in a context which consists of four different objects chosen randomly from the total space of logically possible objects. The speaker randomly chooses an object from this context as the 'topic', subsequently chooses a random feature from this topic as their 'target meaning', and finally produces the signal which is most strongly associated with this meaning in their lexicon (or adds a new signal for this meaning to their lexicon if none of the signals were associated with the target meaning so far). The hearer then first interprets the speaker's utterance (for the purpose of measuring agents' communicative success), and subsequently updates their lexicon by increasing the association weights between the speaker's signal and each of the object features present in the context. However, a distinction is made between the 'physical context', which contains all four objects, and the hearer's 'learning context', which can be narrowed down by means of joint attention. *Checking attention* reduces the number of objects in the context for both interpretation and learning, while *following attention* and *directing attention* do so only for learning. The latter two forms of joint attention are modelled as a 'follow-up' to an initial communicative interaction. (Think of this as the caregiver first introducing a novel word to the child and subsequently directing her attention to another object to which that word also applies; or the child subsequently directing the caregiver's attention to another object to seek confirmation of whether the novel word also applies to that object.)

When the hearer uses *checking attention*, the learning context consists only of the features of the object that the speaker chose as topic (because in checking attention speaker and hearer were both already attending to this object). *Following attention*, as mentioned above, is preceded by a regular interaction in which the speaker randomly chooses a target meaning, produces a signal accordingly, and the hearer interprets this signal based on all objects present in the context. The speaker then directs the hearer's attention to another object in the context which also has the target feature, and the hearer's learning context is reduced to only the features of this object which is now in joint attention. However, if no second object is found in the context which

contains the target feature (because contexts are generated randomly), the learning context is not reduced and remains equal to the physical context. Finally, when using *directing attention*, the hearer — after first interpreting the speaker's signal — directs the speaker's attention to an object in the context which has the feature that the hearer interpreted (rightly or wrongly so) as the speaker's target meaning. The speaker then provides feedback as to whether the target meaning is indeed a feature of this object. If the speaker's feedback is positive, the hearer's learning context is reduced to the object now in joint attention. If the speaker's feedback is negative, the learning context consists of the full context except for that object. Note that the hearer is more likely to receive positive feedback, and therefore to reduce the learning context more significantly, if her interpretation of the speaker's signal was correct. Directing attention thus becomes more useful as the hearer's lexicon takes shape and aligns with that of the speaker.

Kwisthout et al.'s model also allows for the combination of two or all of these forms of joint attention. When multiple joint attention mechanisms are combined, the learning context is simply further reduced to only the features in the cross-section of the learning contexts yielded by the separate mechanisms. Kwisthout et al. explored how each possible combination of none, some or all of these forms of joint attention affect a population's ability to construct an informative lexicon (measured on the basis of communicative success). In each simulation, each agent in the population started out with an empty lexicon, and pairs of agents were chosen at random to interact with each other, with each agent taking turns as both speaker and hearer. As mentioned above, the model did not include any generational turnover, so simulations simply consist of the same population of agents being followed over interactions.

Simulation results showed that whether populations used checking attention or not had the biggest impact on whether they reached maximal communicative success. In all conditions that included checking attention, populations reached 100% communicative success in most simulations, whereas of the conditions without checking attention only those with following attention did so occasionally. In the conditions which included checking attention, the addition of following and/or directing attention reduced the amount of time populations needed to reach maximal communicative success, with the condition including all three forms of joint attention converging the quickest.

These results make sense given that the average size of the learning context was reduced most strongly by checking attention, followed firstly by following attention,

and then by directing attention. The condition with the smallest average learning context size was (unsurprisingly) the condition which combined all three forms of joint attention, but checking attention made the biggest difference in terms of reducing the context size. This is a direct result of how the three forms of joint attention are implemented: checking attention consistently reduces the learning context to one object, while following attention only does so when a second object with the target feature happens to be present, and directing attention relies even more on chance unless the hearer's lexicon has already started to take shape and align with the speaker's.

In sum, the simulation results of Kwisthout et al. (2008) demonstrate firstly that joint attention as a mechanism for reducing referential uncertainty can facilitate not only word learning but also language emergence (in the sense of a group of agents establishing a shared set of form-meaning mappings). Secondly, their results suggests that the order in which the different mechanisms of joint attention develop in children corresponds to the ranking of these mechanisms in terms of how useful they are in facilitating language emergence.

Gong and Shuai (2012) went further to explore not just how joint attention facilitates language emergence, but also how joint attention may have co-evolved with language. They implemented joint attention as the ability to infer a speaker's communicative intention from nonlinguistic information (i.e. without having to know the meaning of the speaker's utterance). During language learning and communication, each utterance received from the agent's cultural parent or communication partner is accompanied by one such environmental cue, and the agent's level of joint attention was implemented as the probability of correctly inferring the speaker's communicative intention from this environmental cue. In contrast to Kwisthout et al.'s (2008) model, joint attention in this model is thus a continuous variable. Languages in Gong and Shuai's model are transmitted culturally through inter-generational communication (and further aligned with those of peers during *intra*-generational communication), but agent's level of joint attention is transmitted genetically.

Gong and Shuai (2012) initialised the first generation of their populations with a mixture of different joint attention levels, and a rudimentary initial signalling system (they report however that initialising populations with no linguistic knowledge whatsoever yielded similar results). Gong and Shuai then compared how the population's levels of language and joint attention evolved under different types of selection on communicative success. Specifically, they compared biological selection (where agents who

are more successful at understanding others have more offspring to whom they transmit their joint attention gene) with cultural selection (where agents who are more successful at understanding others are more likely to be chosen as a cultural parent; i.e. to provide linguistic input for agents in the next generation).

Simulation results showed that when populations are not subjected to any selection pressure for communication, the population's initial average level of joint attention had to exceed a certain threshold in order for the population to further expand the language and thereby increase its communicative success. When biological selection on communicative success was added however, this led to an increase in the average level of joint attention over generations, which in turn allowed all populations (regardless of their initial average level of joint attention) to expand their linguistic system and reach high levels of communicative success. In contrast, when populations were exposed to cultural selection, this did not lead to similar improvements in the populations' languages.

These results makes sense given that agents acquire their language through communication, and that this process relies on the hearer's level of joint attention. Gong and Shuai (2012) implement languages as having both semantic and syntactic structure, such that an agent's lexicon consists of a set of lexical rules (either holistic or compositional) which can generate an utterance given a meaning and vice versa. During language-learning (as well as intra-generational communication), the hearer integrates the communicative intention inferred from the environmental cue with the candidate meaning offered by the linguistic rules in their lexicon. Whenever the interpretation offered by the linguistic rules is incomplete (because the utterance contains elements that the hearer has not encountered before), the hearer supplements their interpretation with the communicative intention inferred from the environmental cue. This means that a learner with a low level of joint attention will very regularly assign incorrect meanings to new utterances, leading to unfaithful cultural transmission of the language. Therefore, if a cultural parent is selected on the basis of having high communicative success, their useful language is unlikely to be transmitted faithfully to a learner who has a low level of joint attention. Thus, cultural selection on languages in this model can only have an effect if combined with a high level of joint attention, or biological selection leading to such a high level of joint attention. In fact, Gong and Shuai (2012) found that when populations were subjected to both biological and cultural selection, the outcome was no different from that under biological selection alone,

indicating that cultural selection caused no added improvements in the population's language on top of what was achieved by the level of joint attention increasing through biological selection.

Given that Gong and Shuai's (2012) model of joint attention is relatively abstract (i.e. defined as the probability of an agent correctly inferring a speaker's communicative intention from nonlinguistic information), this implementation could equally model perspective-taking or mindreading. However, in contrast to the joint inference model presented in Chapter 3, Gong and Shuai's model does not provide a straightforward way in which joint attention might co-develop with language-learning. As a consequence, this model also does not allow for an exploration of how the cultural evolution of language might improve a populations' joint attention abilities over generations, without biological adaptation playing a role.

In the iterated learning model described in Section 4.2 below, agents' ability to infer a speaker's referential intention from the context (i.e. from extralinguistic information) depends on whether they correctly infer their cultural parent's perspective. However, agent's success at inferring their parent's perspective in turn depends on the informativeness of the parent's lexicon (as shown in Chapter 3). In contrast to Gong and Shuai's (2012) model, this model does not include biological adaptation of agents' ability to learn about each other's perspective: every agent enters the population with the same learning abilities and the same prior bias. The only attribute that is transmitted over generations is the lexicon, which is transmitted culturally (through iterated learning). Thus, this model investigates the *cultural* co-evolution of lexicons and perspective-inference. It explores under what circumstances a population of agents can culturally evolve an informative lexicon from scratch, when the faithful transmission of such a lexicon depends on correctly inferring perspectives, and correctly inferring perspectives in turn depends on the lexicon being informative in the first place. This provides a test case for exploring whether culture *itself* can bootstrap culture (Heyes, 2012a, 2018).

4.2 Iterated lexicon learning with perspective-taking agents

4.2.1 Cultural transmission

In the iterated learning version of the developmental model described in Chapter 3, lexicons are transmitted culturally; that is, learned from data produced by the previous

generation. Each agent in generation G_t receives input from a single cultural parent from generation G_{t-1}. All agents in the very first generation of a population are initiated with a completely uninformative lexicon which maps each signal to each referent. Populations thus really have to start from scratch when it comes to evolving informative lexicons. In contrast with agent's lexicons, their perspectives are constrained by the environment. All agents within a given generation share the same perspective, but this alternates over generations such that all learners in generation G_t have the opposite perspective to their cultural parents from generation G_{t-1}. This means that egocentric learners have an unhelpful bias, just like in the simulations described in Chapter 3.

Learners' data consists of a set of contexts combined with corresponding utterances which are produced by their cultural parent. Based on this data, learners try to infer both their parent's perspective and their parent's lexicon (using Bayesian inference as described in Chapter 3), and subsequently select a composite hypothesis (i.e. perspective + lexicon) based on their posterior probability distribution. Hypothesis selection follows the *sampling* method (see Section 4.1.1): the probability of a particular hypothesis being selected is equal to the probability assigned to that hypothesis in the learner's posterior distribution. The learner then assigns the selected perspective to their cultural parent, and adopts the selected lexicon as their own. (As mentioned above, the learner's perspective is innately specified and does not change after learning.) As in Chapter 3, learners only need to consider two possible perspective hypotheses, which are maximally different from each other. Also as in Chapter 3, the learners' hypothesis space of lexicons consists of all logically possible 3x3 lexicons that associate at least one signal with each referent (343 lexicons in total).

4.2.2 Data and the transmission bottleneck

Contexts

In order to speed up simulation run times, the learners in the iterated learning version of the model observe only a fixed set of 'maximally informative' contexts. These are contexts which create maximally distinct saliency distributions for the two possible perspectives. (See Equation 3.1 in Chapter 3 for how a given context and perspective combination give rise to a saliency distribution over potential referents.) Context informativeness is defined as the sum of differences in referent probability ratios between the two different perspectives. The rationale behind this measure is explained in Appendix

A, but to give an example, the maximally informative context $c_1 = [0.1, 0.2, 0.9]$ yields the saliency distributions $[0.50, 0.44, 0.06]$ for perspective $p = 0.0$ and $[0.08, 0.17, 0.75]$ for perspective $p = 1.0$ (see also Figure 3.1 in Chapter 3). These two saliency distributions are very different in terms of the *ratios* between the probabilities of the different referents. (These ratios are independent from how the referents are ordered, and thus do not require knowledge of which signal maps to which referent in order to be observed; although how clearly they will be reflected in the data depends on the informativeness of the underlying lexicon.) For $p = 0.0$, two of the referents are almost equally likely to be chosen as intended referents, whereas the third is very unlikely. For $p = 1.0$ in contrast, the most salient referent is more than four times as likely to be chosen as the second most salient referent, which in turn is about twice as likely to be chosen as the least salient referent. We can contrast this with a less informative context $c_2 = [0.3, 0.4, 0.6]$ which yields the saliency distributions $[0.41, 0.35, 0.24]$ for $p = 0.0$ and $[0.23, 0.31, 0.46]$ for $p = 1.0$. The ratio differences between the two perspectives are much smaller in this case, which will make them harder to distinguish based on the frequencies with which the different referents are selected.

Bottleneck

Every learner's input consists of a dataset of 120 observations, where each observation is composed of a context accompanied by one utterance produced by the cultural parent. This number of observations was chosen on the basis that it is (on average) sufficient for an egocentric learner to reach a posterior probability of $P(\ell) > 0.5$ on the correct lexicon hypothesis, for just over half of the possible input lexicon types (eight out of fourteen). Specifically, this is the case for all lexicon types with an informativeness level of $ca(\ell) > 0.5$, as well as those with $ca(\ell) = 0.43$ and $ca(\ell) = 0.33$ (see also Appendix B). This threshold of $P(\ell) > 0.5$ on the correct lexicon hypothesis is simply a proxy for a posterior distribution in which there is no lexicon hypothesis to which more posterior probability is assigned than the correct lexicon hypothesis. In other words, at this point the learner's belief in the correct lexicon hypothesis has exceeded their belief in all other 'competitor' hypotheses added together. Learners use the sampling method to select their lexicon after learning, so other lexicons still have a chance of being selected after the $P(\ell) > 0.5$ threshold is reached, but the correct lexicon now has a higher probability of selection than any other. In sum, with 120 observations per learner, just

over half of the lexicon types have a reasonable chance of being transmitted faithfully over generations.

However, none of the lexicon types exceed the threshold of $P(\ell) = 0.95$ after 120 observations. Thus, 120 observations (combined with sampling as a method of lexicon selection) instantiates a transmission bottleneck (as discussed in Section 4.1.1), which means that not enough information can be accumulated within this timeframe to perfectly learn the input lexicon (Kirby, 2000). As discussed in Section 4.1.2, such a transmission bottleneck is necessary for processes of convergent transformation and cultural selection to have an effect (i.e. there has to be some chance of the learner transforming the lexicon they received as input, thereby generating new variants). The transmission bottleneck also means that lexicons which require less information to be acquired accurately than others — such as a lexicon which uses only one signal for each referent ($ca = 0.33...$), or a lexicon with one-to-one mappings between signals and referents ($ca = 0.90$) — have a better chance of being transmitted faithfully than others. The amount of observations required to learn a given lexicon type is roughly negatively correlated with its informativeness. As described in Chapter 3, the informativeness of a lexicon is measured as its communicative success 'with itself'. In other words, as the probability that a speaker and listener who both use that lexicon will correctly interpret each other's utterances (see Equation 3.7 in Chapter 3). This measure thus reflects how much information the lexicon provides about the speaker's intended referent, which in turn determines (in part) how many observations are required to learn the lexicon. However, this is only part of the picture, as can be seen in Figure 4.2, which shows the number of observations required to reach the $P(\ell) > 0.5$ threshold, ordered by informativeness category (i.e. lexicon type).

Another factor which influences how many observations are needed to acquire a given lexicon is the amount of signals it makes use of. A set of observed data in which only two of the three possible signals ever occur has a low likelihood under any lexicon hypothesis which uses all three signals. And given that the hypothesis space contains fewer lexicons that use only two signals than lexicons that use all three (75 and 265, respectively), the correct lexicon can then be learned relatively quickly. This advantage of 1-signal and 2-signal lexicons is shown in the bottom panel of Figure 4.2.

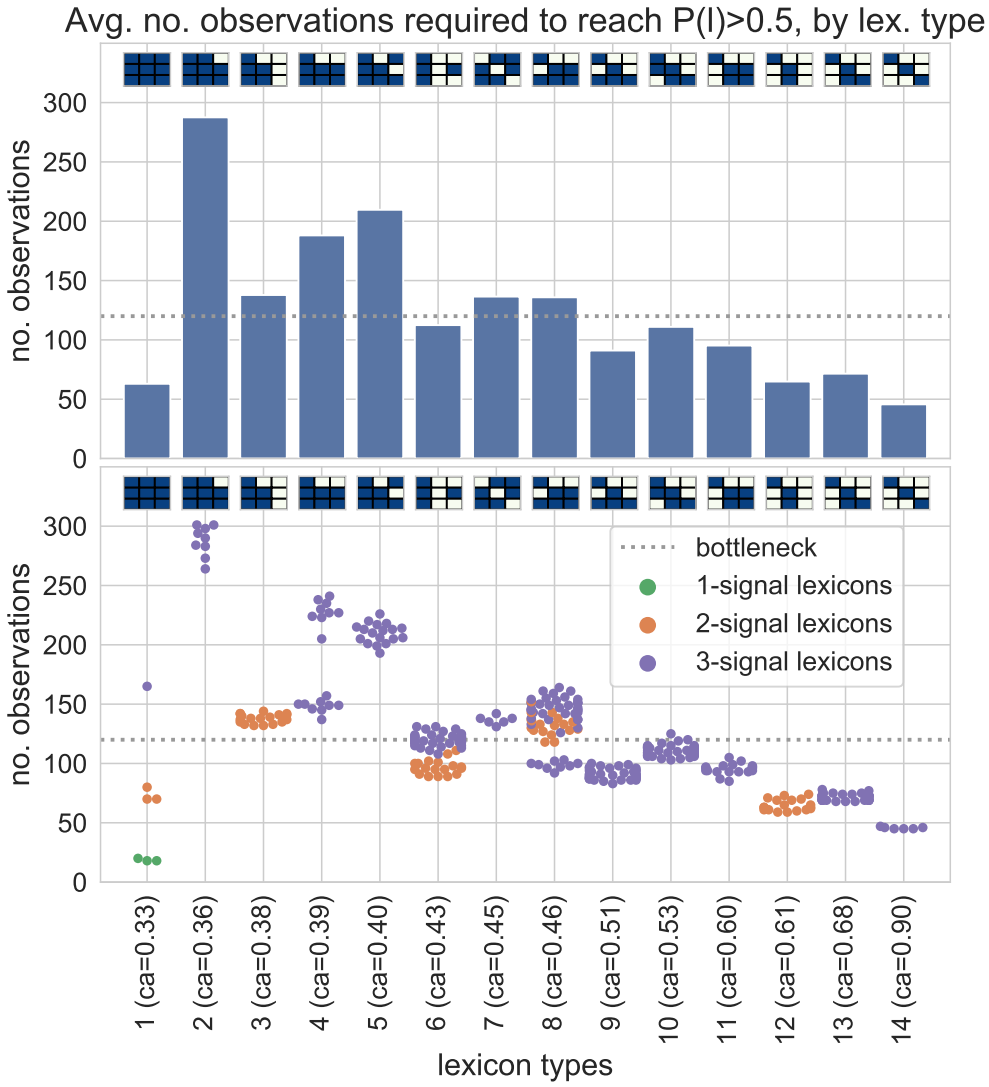


Figure 4.2: Average number of observations required for an egocentric learner to reach a posterior belief of $P(\ell) > 0.5$ in the correct lexicon hypothesis when observing only optimal contexts, categorised by informativeness of input lexicon. Top panel shows means over all lexicons within a given lexicon type, where each individual lexicon's value is in turn a mean over 100 independent simulation runs. Bottom panel shows individual lexicons (same means over 100 simulation runs), colour-coded by how many signals they make use of. Grey dashed line shows number of observations used for all simulations reported below (i.e. the transmission bottleneck; 120 observations). Matrices at top of panels represent example lexicons for each corresponding informativeness category (i.e. lexicon type), with referents on the rows and signals on the columns. Blue squares in these matrices represent an association between the corresponding referent and signal, while white squares represent the absence of such an association. Note that these matrices show only a single example lexicon for each informativeness category, chosen for illustrative purposes.

Finally, as discussed in Chapter 3 (Section 3.5), the fact that the lexicon type with the lowest informativeness level (i.e. type 1, with $ca = 0.33...$) takes so few observations to learn (see Figure 4.2) is partly due to the fact that this is the only lexicon type which does not require knowing the speaker's perspective in order to be learned accurately. This is a result of the fact that all lexicons within this lexicon type use only signals that are associated with every single referent. Therefore, which referent the speaker chooses to communicate about does not alter the probabilities with which they will use the different signals, and thus the context and the speaker's perspective have no influence on their utterances. In Bayesian terms, for all lexicons comprised in this lexicon type, the likelihood of the data is equally high given the correct perspective hypothesis as it is given the incorrect perspective hypothesis.

In sum, although the amount of observations required to learn a given lexicon type is roughly negatively correlated with its informativeness, there are other factors which influence a lexicons learnability as well, meaning the transmission bottleneck does not straightforwardly select for more informative lexicons.

4.2.3 Selection

All simulation results presented below were obtained with populations of 100 agents, where each new generation G_{t+1} is formed by replacing all agents of G_t at once. As described above, every new agent receives data from a single cultural parent. Three different selection conditions are explored below, which determine the probability with which agents from G_t are chosen as cultural parents for agents of G_{t+1}. In all cases, agents are sampled with replacement from this probability distribution; that is, a single cultural parent can have multiple learners.

In the *No Selection* condition, agents from G_t are sampled with uniform probability. In the *Selection on lexicon-learning* condition, agents' probability of being chosen as parent is proportional to the posterior probability they assign to the correct hypothesis about the lexicon of their own cultural parent from generation G_{t-1}. (I.e., the better an agent has learned their parent's lexicon, the more likely they are to be chosen as a cultural parent themselves.) The *Selection on perspective-inference* condition works similarly, but instead the decisive factor here is how well the agent has learned their parent's perspective. Thus, agents' probability of becoming a cultural parent is proportional to the posterior probability they assign to the correct hypothesis about their

own cultural parent's perspective.

Finally, in the *Selection for communication* condition, the probability that an agent is chosen as a cultural parent is determined by their communicative success with their own cultural parent. Because by this point the agent from G_t has adopted both a lexicon of their own and assigned a perspective to their cultural parent (be it correct or incorrect), both these types of knowledge are integrated in the measure of communicative success. The learner agent a_l takes on the role of listener, and uses Bayes' rule to derive the probability that their cultural parent's referential intention was r given that the parent used signal s, essentially 'inverting' the model of how the cultural parent generates utterances given a context, as shown in Equation 4.1. However, the learner uses their own lexicon ℓ_{a_l} and their model of their cultural parent's perspective $p'_{a_{cp}}$ in this interpretation procedure, either or both of which may not correspond to the reality about their cultural parent.

$$P_{a_l}(r \mid s,\, \ell_{a_l},\, c,\, p'_{a_{cp}}) \propto P_{a_{cp}}(s \mid r,\, \ell_{a_l})\, P(r \mid c, p'_{a_{cp}}) \tag{4.1}$$

The communicative success between a cultural parent a_{cp} and their learner a_l in a context c is defined as the average probability that a_{cp} will produce a signal which enables a_l to correctly identify a_{cp}'s intended referent, over all possible referents, as shown in Equation 4.2.

$$cs(a_{cp}, a_\ell \mid c) = \sum_{r \in R} \sum_{s \in S} P(s \mid r, \ell_{a_{cp}}) P(r \mid c, p_{a_{cp}}) \cdot P(r \mid s, \ell_{a_l}) P(r \mid c, p'_{a_{cp}}) \tag{4.2}$$

where R stands for the full set of potential referents, $\mathcal{S}$ for the full set of signals, $\ell_{a_{cp}}$ for the lexicon of the parent, ℓ_{a_l} for the lexicon of the learner, $p_{a_{cp}}$ for the parent's perspective, and $p'_{a_{cp}}$ for the learner's model of their cultural parent's perspective (which may be correct or incorrect). The full communicative success of a learner a_l is not measured over a single context, but the average is taken over a set of six randomly generated contexts, as shown in Equation 4.3.

$$CS(a_{cp}, a_l) = \frac{1}{|C|} \sum_{c \in C} cs(a_{cp}, a_l \mid c) \tag{4.3}$$

where C stands for the full set of 'test' contexts. Note that these contexts are randomly generated, so not the same as the fixed set of maximally informative contexts

that the learner was trained on.

Because the learner's knowledge about their cultural parent's perspective is used in this measure of communicative success, the measure takes into account only the learner's comprehension success, not their success at producing signals that their cultural parent will understand. This also keeps the *Selection for communication* condition as similar as possible to the *Selection on lexicon-learning* and *Selection on perspective-inference* conditions, in the sense that all that matters for the learner's probability of being selected is the learner's knowledge about their cultural parent, not the cultural parent's knowledge about the learner.

The three selection pressures described above can simulate either biological selection (where more successful agents have more offspring) or cultural selection (where more successful agents are more likely to be chosen as a model to learn from by other agents; Boyd and Richerson, 1985). Biological selection is used here simply to mean that selection is driven by the environment rather than by the offspring. The model does not include any form of genetic inheritance; the only trait that is transmitted is the lexicon, which is transmitted culturally. However, this model remains agnostic as to the mechanism through which a given selection pressure is exerted; what matters here is which trait (lexicon-learning, perspective-inference or communication) is selected on.

4.3 Emergence of informative lexicons in populations

This section explores under what circumstances a population of agents who start out without any linguistic conventions whatsoever are able to establish informative lexicons (i.e. every agent in the first generation starts with the same completely uninformative lexicon, which associates each signal with each referent). As discussed in Chapter 3, the developmental results obtained with this model showed that learners can infer a speaker's lexicon and perspective successfully only if the speaker uses a lexicon that is at least somewhat informative. The question posed in the current chapter is how a population of agents who develop in this way (where lexicon-learning requires correct perspective-inference, but correct perspective-inference in turn requires an informative lexicon) can build such an informative lexicon from scratch. In order to answer this question, the learning model presented in Chapter 3 is embedded in an iterated learning model as described in the previous section. To recap, the learning model is that of a Bayesian learner who has to simultaneously infer a speaker's lexicon and perspective

based on observing that speaker's utterances in context. I will discuss the iterated learning results obtained with this model in two parts. In Section 4.3.1 below I will describe how the lexicons in the population change over generations under four different selection conditions (*No selection*, *Selection on lexicon-learning*, *Selection for communication* and *Selection on perspective-inference*), and how this affects agents' success at communicating and inferring each others' perspectives. In Section 4.3.2 I will discuss how these different selection pressures interact with the learners' egocentric perspective bias.

4.3.1 Emergence of informative lexicons under different selection pressures

In all simulations reported below, the first generation of agents in the population starts out sharing the same, completely ambiguous lexicon, in which every signal is associated with every referent. Given that the number of referents in these simulations is set to three, this initial lexicon has an informativeness of 0.33... *ca.* (Informativeness was defined in Chapter 3 as the average communicative accuracy of a lexicon 'with itself'; that is, between two agents who both use that lexicon, see Equation 3.7.)

Informative lexicons emerge under all three selection pressures

Figure 4.3 shows how the average informativeness of the lexicons in the population changes over generations in the four different selection conditions, and compares populations of unbiased agents and populations of egocentric agents. Egocentric agents have a strong prior bias in favour of the hypothesis that other agents share their perspective on the world. In all simulations reported in this thesis, this egocentric bias is unhelpful, because learners in fact receive input from a speaker who has a perspective exactly opposite to theirs.

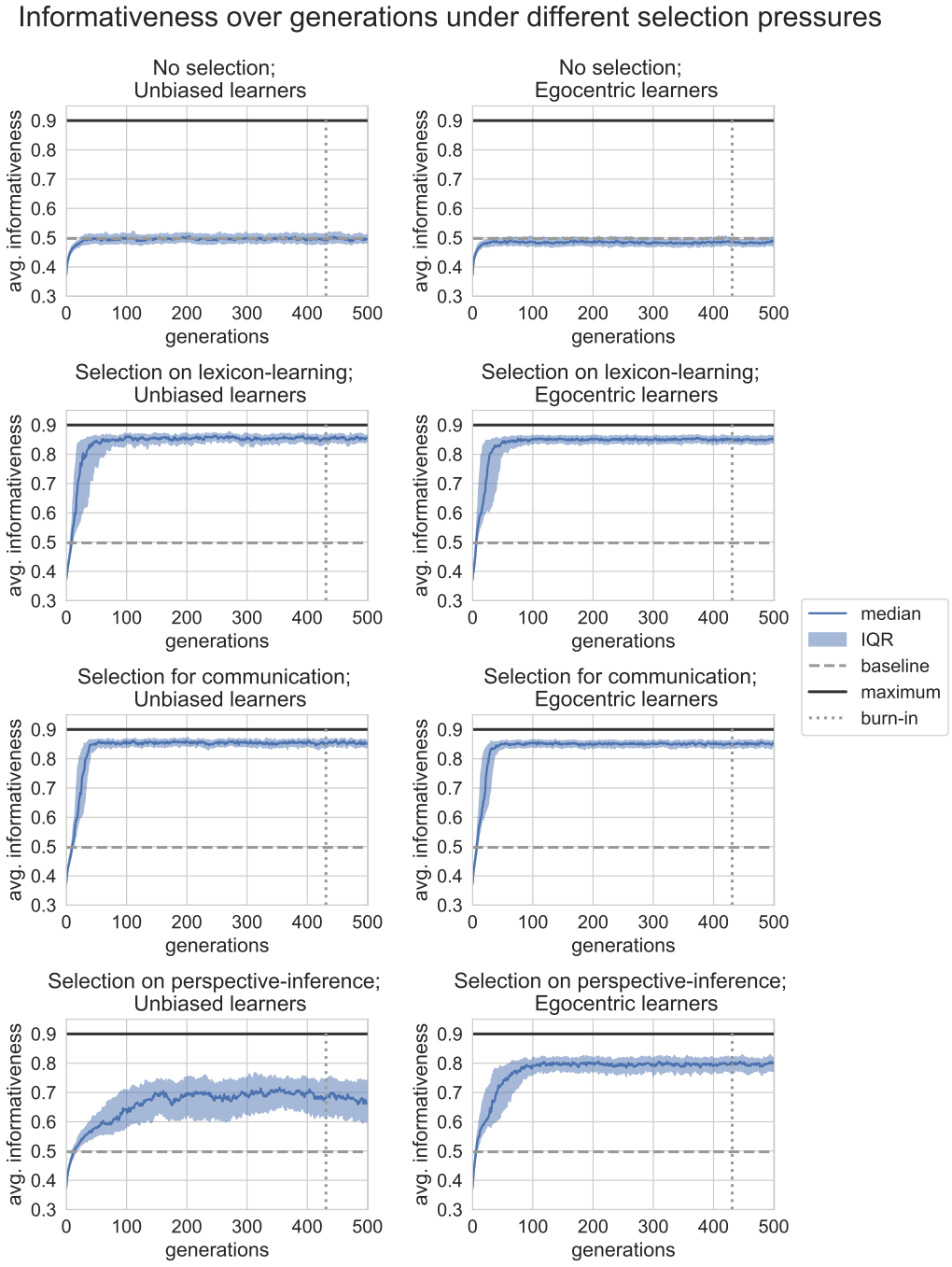


Figure 4.3: Average informativeness of lexicons in population over generations. Solid blue line shows median and shaded area shows interquartile range over 100 independent simulation runs. Dashed grey line shows the baseline informativeness one would expect if agents are picking lexicons at random. Minimum informativeness is 0.33... *ca* (given lexicon size 3x3), which is the starting point for all populations. Maximum informativeness is 0.90 *ca* (given production error $\epsilon = 0.05$), which is indicated by the solid black line. Dotted grey line indicates the final generation of the burn-in period which is discarded for calculating the populations' success and the equilibrium distribution over lexicon types (both reported below).

The results in Figure 4.3 show that in the *No selection* condition, the average informativeness quickly reaches the level that would be expected if agents are selecting lexicons at random. Under each of the three selection pressures in contrast, the average informativeness of the lexicons used by the population increases rapidly and reaches a stable high level after about 100 generations. This effect is strongest in the *Selection on lexicon-learning* and the *Selection for communication* conditions. In most conditions, the average informativeness rises to nearly the same level in populations of unbiased learners and populations of egocentric learners, but not quite: informativeness in the egocentric populations hovers just below that of the unbiased populations. However, this difference flips and becomes more pronounced in the *Selection on perspective-inference* condition. Surprisingly, there the average informativeness increases *more* in populations of agents who have the unhelpful egocentric bias than it does in unbiased populations, and unbiased populations also show more variability. The interaction between agents' prior bias and the different selection pressures is discussed further in Section 4.3.2 below. Exactly which lexicon types are present in the population in these different conditions is discussed in more detail below (see Figure 4.5).

Figure 4.3 also depicts the 'burn-in' period which was discarded in measures (reported below) that are concerned with characteristics of the populations after they have reached convergence. This 'convergence point' was defined as the generation from which onwards the variation in average informativeness remains within a range of 0.1 *ca* for a minimum of 50 consecutive generations, in each individual simulation run. Different conditions show a wide range of different convergence points (ranging from 62 generations in the *Selection for communication* condition with egocentric agents to 431 generations in the *Selection on perspective-inference* condition with unbiased agents). However, in order to keep the number of generations used to compute further measures constant across conditions, the highest of these convergence points (431 generations) was used uniformly across conditions as a burn-in period.

The emergence of informative lexicons leads to both more successful communication and more successful perspective-inference

Figure 4.4 shows populations' average success at communicating and inferring each others' perspectives after convergence. In order to speed up simulation run times, these success measures are calculated between each learner of the generation of in-

terest and their respective cultural parent (for whom the learner has already inferred a perspective) rather than between learners and their peers in the same generation. Communicative success is measured as the probability that the learner will correctly interpret the utterances of their cultural parent in a given context, averaged over all possible referents (see equations 4.2 and 4.3 in the previous section). This is the same measure on the basis of which agents are subsequently selected to become a cultural parent for the next generation or not, but evaluated on an independent set of six randomly generated contexts. Perspective-inference success is measured simply as 1 if the perspective hypothesis that the learner selected for their parent is correct, and 0 otherwise.

As Figure 4.4 shows, an increase in average informativeness of the lexicons in the population (see Figure 4.3) causes an increase not just in communication success but also in agents' success at inferring each others' perspectives. Firstly, in the *No selection* condition, both communicative and perspective-inference success are somewhat above chance level. Chance level performance is what would be expected if the lexicons present in the population do not provide any information about speakers' intended referents (i.e. have an informativeness level of 0.33... *ca*). Although this is the lexicon type that populations start out with, Figure 4.3 shows that the average informativeness in the *No selection* condition quickly rises to the level that corresponds to agents picking lexicons at random. That means that after convergence, most agents receive input from lexicons that give them at least some information about the speaker's intended referents. This in turn allows those agents to interpret their cultural parent's utterances correctly somewhat more often than expected by chance, and also gives them a way into learning something about their parent's perspective.

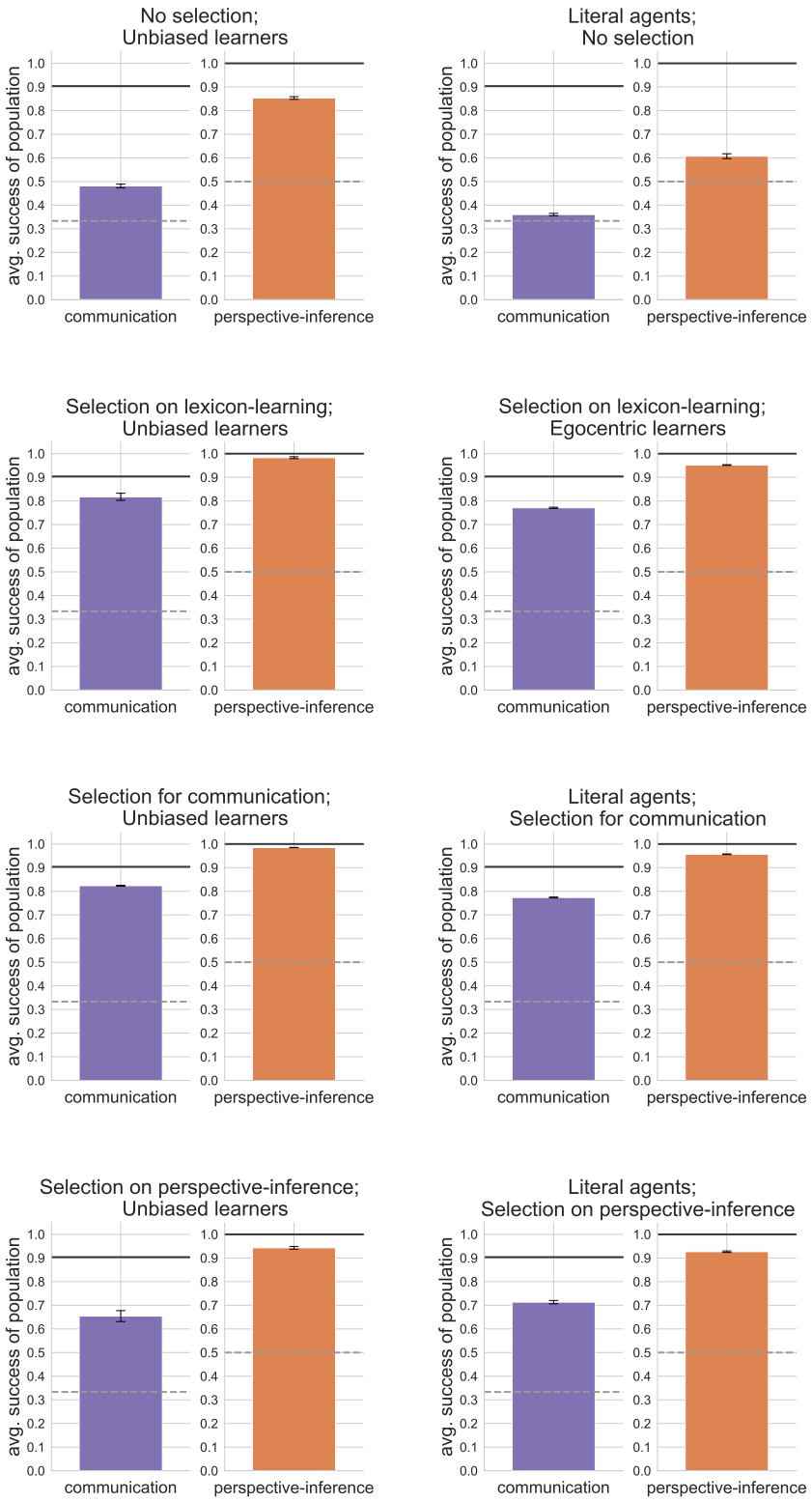


Figure 4.4: Average success of the population at communicating with their cultural parent and inferring the perspective of their cultural parent (calculated independently from agents' fitness), after convergence (i.e. measured from generations 431 to 500). Graph shows grand means and 95% confidence intervals over 100 runs, 69 generations and 100 agents (i.e. each bar shows the grand mean success of 690,000 individual agents). Dashed grey line indicates chance level and solid black line indicates ceiling.

Secondly, agents' success increases most strongly in the *Selection on lexicon-learning* and *Selection for communication* conditions, again in line with the levels of informativeness that involves in these conditions. As with informativeness, populations' success reaches slightly higher levels for the unbiased learners than it does for the egocentric learners in the *No selection*, *Selection on lexicon-learning* and *Selection for communication* conditions. Finally, in the *Selection on perspective-inference* condition, agents' communication success also reaches a level in line with the average informativeness that evolves in this condition. Just as with informativeness, this success measure shows the opposite pattern to the other selection conditions: communication success increases more in the egocentric populations than it does in the unbiased populations. Agents' perspective-inference success does not show this flipped pattern in line with the informativeness measure, but this is explained by how the two different success measures interact with agents' prior bias. Agents' communication success is not affected to the same extent by the egocentric bias as their perspective-inference success. Agents' communication success depends on three different factors; although agents do make use of their knowledge of their parent's perspective to interpret utterances, communication success depends first and foremost on the informativeness of the agents' lexicons and the extent to which those lexicons are aligned. Agents' perspective-inference does depend on the informativeness of the parent's lexicon as well, but unbiased learners have a headstart of 0.4 prior probability on the correct perspective hypothesis compared to egocentric learners.

All selection pressures lead to convergence on the most informative lexicon type

To gain more insight into which lexicons cause the increase in informativeness shown in Figure 4.3, Figure 4.5 shows the proportions with which the different lexicon types are selected by agents after convergence.

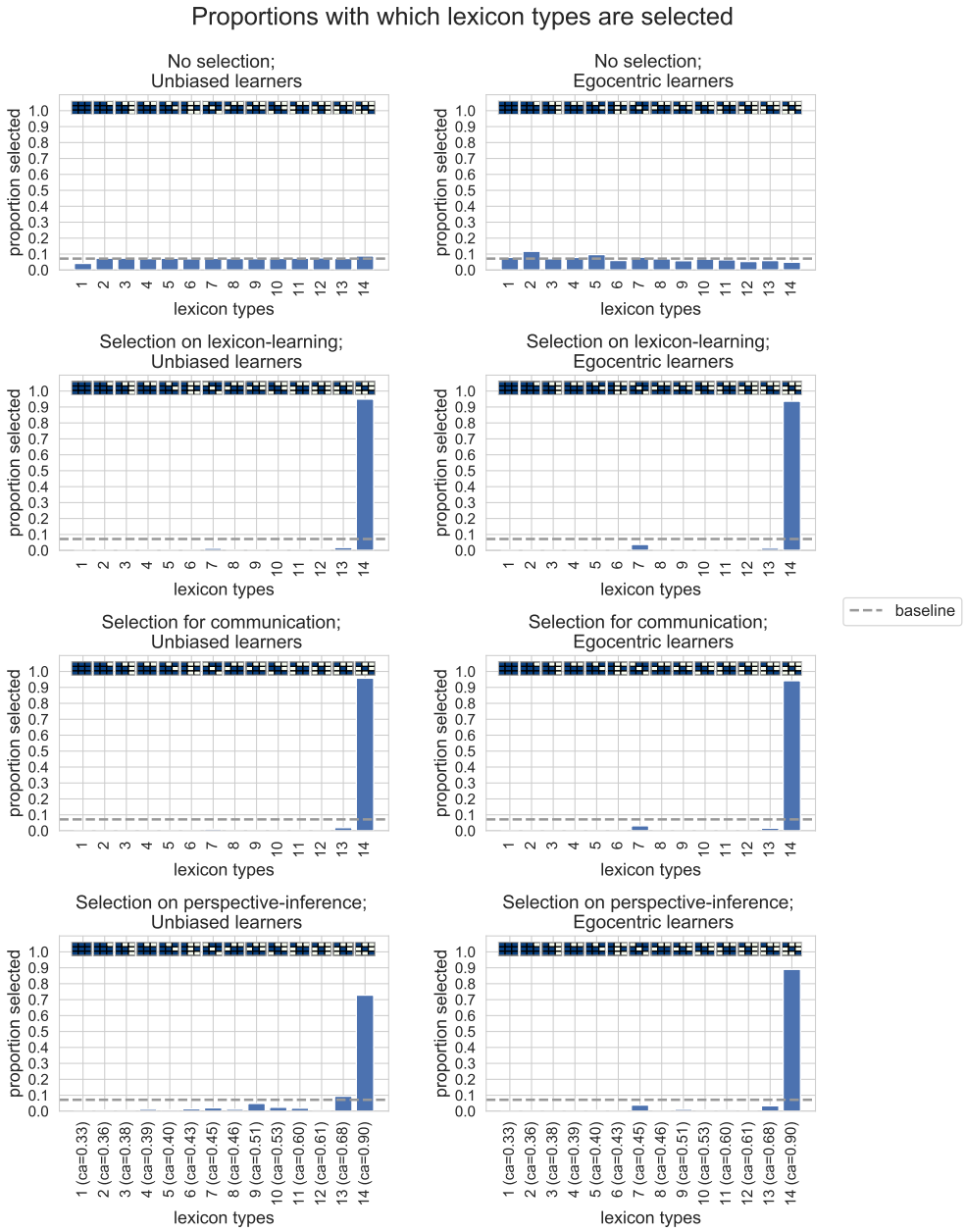


Figure 4.5: Average proportions with which agents select the different lexicon types after convergence (i.e. measured over generations 431 to 500). Graphs show grand means over 100 runs, 69 generations of 100 agents, and the total set of possible lexicons within each informativeness category (ranging between 6 and 63). Each subplot thus summarises 690,000 lexicon-selection events in total; which is equal to 2,012 times the number of lexicon hypotheses in the learners' hypothesis space. The x-axis shows all informativeness categories that exist for 3x3 lexicons, with informativeness levels ranging from lowest possible (0.33... *ca*) to highest possible (0.90 *ca*), given error rate $\epsilon = 0.05$. Dashed grey line shows the baseline distribution over lexicon types one would expect if agents select lexicons at random.

Firstly, Figure 4.5 shows that the *No selection* condition leads to a nearly uniform distribution over lexicon types. However, the unbiased populations under this pressure

show a slight underrepresentation of the least informative lexicon type (type 1) and a slight overrepresentation of the most informative lexicon type (type 14). The egocentric populations show the opposite pattern: several of the less informative lexicon types are overrepresented (especially types 2 and 7), while the more informative lexicon types are slightly underrepresented. In each of the selection conditions, the dominant lexicon type after convergence is the most informative one (type 14), which comprises all lexicons that consist of only one-to-one mappings between referents and signals. In most of the conditions with selection, the only two lexicon types co-existing with type 14 after convergence are type 7 ($ca = 0.45$) and type 13 ($ca = 0.68$). The only exception to this is the *Selection on perspective-inference* condition with unbiased learners, in which several other lexicon types continue to have some representation after convergence as well. This is in line with the higher degree of variability in the average informativeness in this condition, as shown in Figure 4.3.

However, the finding that the average informativeness, and therefore also populations' success, increases in the *Selection on lexicon-learning* condition is more sensitive to the size of the lexicon and the width of the bottleneck than the finding that average informativeness increases in the *Selection for communication* and *Selection on perspective-inference* conditions. Which lexicons are transmitted over generations in the *Selection on lexicon-learning* condition depends for the most part on how faithfully each lexicon type is transmitted, because faithful learning is what determines agents' success in this condition. The most faithfully transmitted lexicon types are always the least informative lexicon type (because learning its mappings correctly does not require knowledge of the cultural parent's perspective) and the most informative lexicon type (because it has the least ambiguous mappings of all lexicon types). However, which of these two types takes fewer observations to learn, and therefore which of these two types is transmitted most faithfully, depends on a range of parameter settings.

Firstly, in the initial stages of learning, learners accumulate belief in the correct lexicon hypothesis more quickly when given input from the least informative lexicon type than when given input from the most important lexicon type (see Appendix B for the corresponding figures). After a number of observations, the rate of learning for the two different lexicon types crosses over. Therefore, if the bottleneck becomes narrower, the least informative lexicon type will be selected for under a pressure for correct lexicon-learning. Secondly, this cross-over point happens later when the lexicon size is bigger and there are therefore more mappings to learn. Thus, if the lexicon size

increases, the *Selection on lexicon-learning* condition will also increase the chance of populations converging on the least informative lexicon type. Thirdly, the cross-over point between the least informative and most informative lexicon type is also delayed when learners observe randomly generated contexts rather than maximally informative contexts. And finally, the cross-over point is also delayed when learners have an egocentric bias. In sum, whenever parameter settings change in a direction that is more like lexicon transmission in the real world, the *Selection on lexicon-learning* condition will select for the least informative lexicon type rather than the most informative lexicon type as in the parameter settings used in this chapter.

Non-convergence to the prior

The fact that the equilibrium distribution over lexicons in the *No selection* condition is not exactly equal to the learners' prior bias over lexicons (which is uniform) is surprising given that agents select a lexicon hypothesis by sampling from their posterior probability distribution. As discussed in Section 4.1.1, Griffiths and Kalish (2007) demonstrated analytically that in iterated learning models with Bayesian learners who use sampling to select a hypothesis, the ultimate equilibrium distribution over hypotheses (i.e. the stationary distribution) always converges to the learners' prior. The fact that the distributions for the *No selection* condition shown in Figure 4.5 deviate slightly from a uniform distribution may therefore be a result of simulations not having run for sufficiently many generations in order for the effect of the populations' initial lexicon to wash out. However, a numerical calculation of the stationary distributions in the *No selection* condition for 2x2 lexicons using the method described by Griffiths and Kalish, shows that deviation from the prior is likely to be a systematic result for populations of egocentric learners: see Figure 4.6.

In short, the method for numerically calculating the stationary distribution of an iterated learning chain of Bayesian agents as described by Griffiths and Kalish (2007), conceives of such a chain as a Markov chain on hypotheses and data. This allows us to numerically calculate the transition matrix specifying the probability that a learner infers lexicon ℓ_i after receiving data from a speaker with lexicon ℓ_j, for each possible combination of ℓ_i and ℓ_j. This is done by enumerating all possible datasets, and for each dataset calculating the probability that a learner will infer lexicon ℓ_i given the dataset (i.e. the probability of inferring lexicon ℓ_i given the learning algorithm LA and

dataset d: $P_{LA}(\ell_i \mid d)$) and multiplying this with the probability with which an agent with lexicon ℓ_j would produce that dataset (i.e. the probability of dataset d given the production algorithm PA and lexicon ℓ_j: $P_{PA}(d \mid \ell_j)$), . Summing this product of learning and production probabilities over all possible datasets yields the total probability of lexicon ℓ_i being transitioned into from lexicon ℓ_j when generation t learns from generation $t-1$: $q_{ij} = \sum_{d \in D} P_{LA}(\ell_t = \ell_i \mid d) P_{PA}(d \mid \ell_{t-1} = \ell_j)$. Once the full transition matrix is calculated in this way, linear algebra provides that the stationary distribution of this transition matrix is equal to its first eigenvector. Unfortunately this numerical solution becomes intractable for lexicon sizes larger than 2x2 combined with the large number of observations required in the current model, because observations in this model consist of combinations of a context and an utterance, rather than utterances alone. This means that a numerical calculation of the transition matrix requires enumeration of not just all possible datasets, but also all of their permutations.[6] This is due to the fact that which utterance goes with which context matters for calculating both $P_{PA}(d|\ell_j)$ and $P_{LA}(\ell_i|d)$ (given that d consists of combinations of a context c and a signal s), and therefore the order of utterances in the dataset matters.

[6]The number of possible datasets for a given dataset size (i.e. number of observations) $|d|$, and a given number of possible signals $|S|$ is given by $\frac{(|d|+(|S|-1))!}{|d|!\cdot(|S|-1)!}$, which equals 7,381 for 120 observations and three signals. The number of possible permutations for a dataset d equals $\frac{|d|!}{\prod_{|s_i| \in d} |s_i|!}$ where $|s_i| \in d$ stands for the frequency of signal s_i in dataset d. The number of permutations for all possible datasets given 120 observations and three signals ranges from 1 to 1.23×10^{55}.

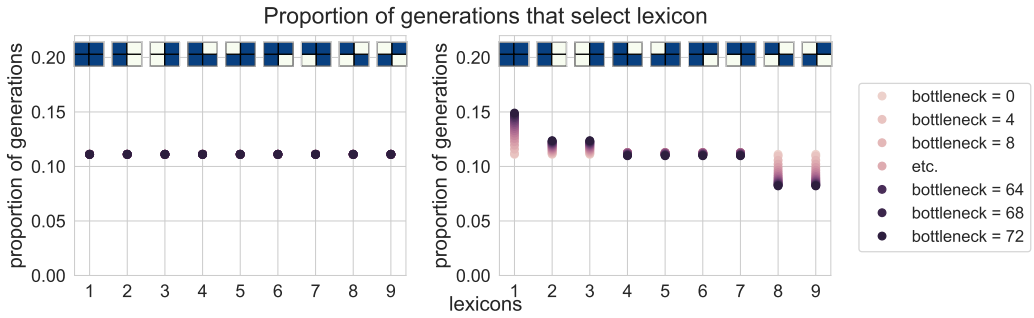


Figure 4.6: Stationary distributions for unbiased learners and egocentric learners in the *No selection* condition, numerically calculated using the method described by Griffiths and Kalish (2007), for lexicon size 2x2 and different bottleneck sizes. Graphs show the full set of possible lexicons (9 in total for 2x2 lexicons) rather than categorising lexicons by informativeness level. Grouping these lexicons into informativeness categories would yield three types: (i) lexicons 1-3 are completely ambiguous, yielding an informativeness level of 0.50 *ca* given two referents; (ii) lexicons 4-7 have an informativeness level of 0.60 *ca*; and (iii) lexicons 8 and 9 are maximally informative given $\epsilon = 0.05$, with an informativeness level of 0.90 *ca*. The maximum bottleneck width of 72 observations shown in this figure is the 2x2 lexicon equivalent of a bottleneck width of 120 observations for 3x3 lexicons. That is, 72 observations is sufficient for learners of 2x2 lexicons to exceed the threshold of $P(\ell) > 0.5$ posterior probability on the correct lexicon hypothesis for more than half of the lexicon types (i.e. types (ii) and (iii); lexicons 4 to 9).

The deviation from the prior distribution over lexicons found for egocentric learners and 2x2 lexicons in Figure 4.6 corresponds to that seen in the *No selection* condition with egocentric learners and 3x3 lexicons (Figure 4.5) in the sense that in both cases the less informative lexicons are overrepresented and the more informative lexicons are underrepresented. The numerically calculated stationary distribution for unbiased learners and 2x2 lexicons on the other hand does not show any deviation from the uniform prior distribution over lexicons. Recall that both unbiased and egocentric learners have a uniform prior distribution over lexicons, and that thus the only difference between the two conditions shown in Figure 4.6 is the absence or presence (respectively) of an egocentric perspective bias. Thus, these results suggest that the presence of this bias increases the probability of learners selecting one of the less informative lexicons. Figure 4.6 also reveals that this effect of the egocentric bias on the stationary distribution over lexicons increases with the width of the bottleneck. This suggests that what may start out as a very small bias of individual egocentric learners towards selecting less informative lexicons can be amplified over generations when the bottleneck is wide and these lexicons thus have a reasonable chance of being transmitted faithfully to the next generation.

The only difference between the current model and the iterated learning chains with Bayesian agents as analysed by Griffiths and Kalish (2007) is the fact that learners in the current model are faced with a joint inference task. As a consequence, learners can have a uniform prior distribution over lexicons, while at the same time having a non-uniform prior over their full hypothesis space. As described in Chapter 3 (Section 3.4), this full hypothesis space consists of every possible combination of lexicon and perspective hypotheses. Thus, although all learners have a uniform prior distribution over lexicons, only the unbiased learners have a uniform prior distribution over the full hypothesis space. This is not the case for egocentric learners, whose prior distribution is strongly biased in favour of their own perspective. As described in Chapter 3, this causes egocentric learners to require more observations to reach the same level of posterior belief in the correct lexicon hypothesis as unbiased learners (see also Appendix B). Although subtle, this causes egocentric learners to have a tendency to underestimate the informativeness of the lexicon they are receiving input from, which is most pronounced for the four most informative lexicon types (see Appendix C). Such a systematic underestimation of lexicon informativeness compared to unbiased learners can cause egocentric learners to be slightly more likely to infer a less informative lexicon given the same dataset, ultimately causing a deviation from the prior distribution over lexicons in the stationary distribution. In sum, the joint inference task that learners in the current model are faced with causes the less informative lexicons to be systematically overrepresented in populations of egocentric learners when no selection pressure is present.

4.3.2 The interaction between learning bias and selection

As discussed in Section 4.1.2, the cultural evolution of a given trait can be directed by two different processes: cognitive biases that are shared between individuals and thereby cause convergent transformations on the one hand, and selection on the other hand. These two processes can be at work simultaneously, and when they are, the directions of their effects can coincide or diverge. The latter is the case for the current model. As described above, learners with an egocentric bias have a slight tendency to underestimate the informativeness of the lexicon they receive input from, and this slight bias can get somewhat amplified over generations such that in the *No selection* condition, the less informative lexicon types are overrepresented in egocentric popu-

lations after convergence. Each of the selection conditions on the other hand biases transmission in the other direction for the parameter settings used here: agents who receive input from a more informative lexicon type are more successful and therefore have a higher chance of being selected to produce data for the next generation.[7]

Figures 4.3 and 4.5 showed that in the *Selection on lexicon-learning* and the *Selection for communication* conditions, the effect of selection in terms of driving up informativeness is almost equally strong in populations of unbiased and populations of egocentric learners. In the *Selection on perspective-inference* condition in contrast, the effect of selection is stronger in egocentric populations than it is in unbiased populations. Thus, an egocentric perspective bias causes selection on perspective-inference to have a stronger effect in terms of increasing the average informativeness of the lexicons in the population, even though the convergent transformations caused by this bias work in the opposite direction (towards lower informativeness). In the current model, this effect makes intuitive sense: the informativeness of the lexicons in the population increases under selection on perspective-inference because agents who receive input from a more informative lexicon are more likely to correctly infer the perspective of their cultural parent. For egocentric learners, the informativeness of the lexicon they receive input from makes a bigger difference to how likely they are to correctly infer their parent's perspective than it does for unbiased learners. For example, given the bottleneck width used here, unbiased learners can reach near-ceiling perspective-learning ($P(p_{cp}) > 0.9$) given any of the five most informative lexicon types as input, while egocentric learners can do so only when given input from the single most informative lexicon type (see Appendix B for the corresponding figure). This implies that the same difference in informativeness between the lexicons of two agents will cause a bigger difference in their 'fitness' under selection for perspective-inference when they are egocentric compared to when they are unbiased.

To explore how such differential 'fitness' differs between conditions, Figure 4.7 shows how the range of different fitness values (defined as an agent's probability of being chosen as cultural parent) changes over the first 50 generations in the different

[7]Note that for the *Selection on lexicon-learning* condition, this effect holds only because under the parameter settings used for the simulations in this chapter the most informative lexicon type has a higher chance of being transmitted faithfully than the least informative lexicon type. As discussed in Section 4.3.1, this effect flips when the bottleneck width decreases, the lexicon size increasing, and/or contexts are generated randomly rather than from a fixed set of maximally informative contexts. Under these circumstances, the *Selection on lexicon-learning* condition will instead favour the least informative lexicon type.

bias*selection conditions. Figure 4.7 shows the difference between the fitness of the most fit agent and the fitness of the least fit agent (the 'maximum differential fitness') as a measure of fitness variation.[8] This figure reveals that fitness reaches higher levels of variation in egocentric populations compared to unbiased populations under all three selection pressures. However, this difference is most pronounced in the *Selection on perspective-inference* condition, followed firstly by the *Selection on lexicon-learning* condition, and then by the *Selection for communication* condition. (This ranking between the *Selection on perspective-inference* condition and the *Selection lexicon-learning* condition is in line with the results of individual learners shown in Appendix B.) Both these selection pressures select directly on how faithfully a given attribute of the cultural parent is learned: recall that agents' fitness in these conditions is defined as proportional to the amount of posterior probability assigned to the correct perspective or lexicon hypothesis, respectively. However, the results of individual learning (see in Appendix B) show that the difference in how important it is to receive input from an informative lexicon for egocentric compared to unbiased learners is greater for learning about a speaker's perspective than it is for learning about a speaker's lexicon. (Compare the extent to which learning curves 'fan out' more for egocentric learners as a function of informativeness than they do for unbiased learners, between perspective-learning in Figure B.3 and lexicon-learning in figure B.2). Thus, although both selection on perspective-inference and selection on lexicon-learning select directly on how well the respective attribute of the cultural parent is learned, the addition of an egocentric bias changes the fitness differential between two lexicon types more under selection on perspective-inference than it does under selection on lexicon-learning.

The difference in fitness variation between unbiased and egocentric populations is smallest in the *Selection for communication* condition. Under this selection pressure, agents are not selected directly on the basis of how much they learned about a given attribute of their cultural parent, but on the basis of how successful they are at correctly interpreting the utterances of their cultural parent. Thus, agents' fitness in this condition depends in part on how well they have learned the lexicon and perspective of their cultural parent (because both are used in interpretation), but not solely. That is, if an agent has misinferred the lexicon of their cultural parent but the two lexicons are still relatively compatible (e.g. sharing the same signal-referent mappings for two of

[8] Using the interquartile range of fitness values or standard deviation from the mean as measures of fitness variation yield similar results.

the signals, but having a slightly different mappings for the third), the agent can still be relatively successful at interpreting their parent's utterances, and therefore have a relatively high fitness.

As Figure 4.7 shows, fitness variation decreases rapidly under selection, and does so more rapidly the higher the peak of fitness variation that is reached. This is not surprising given that selection works by virtue of differential fitness being present in the population. After about 40 generations, each of the selection pressures have eliminated the lexicons that lead to low levels of fitness from the population, and the fitness variation remains at a stable low level (indicating relatively homogeneous populations). At this point, the difference in fitness variation between unbiased and egocentric populations has virtually disappeared.

As discussed in Section 4.1.2, Claidière et al. (2018) found an interaction between convergent transformation and selection similar to the one observed here, in a series of simulations using empirical data from an iterated learning experiment with baboons. That is, they found that convergent transformations in the opposite direction of selection strengthens the effect of selection, relative to when selection is combined with random transformations. Claidière et al. showed that this increased selection strength was due to the convergent transformations first eliminating variation from the population by transforming variants towards the attractor, and subsequently making transmission more faithful because once a variant is in the vicinity of an attractor, it is unlikely to be transformed away from it. This strengthens the effect of selection because selection relies on low levels of transformation. Thus, convergent transformations in this study make selection more effective by means of increasing transmission fidelity. It is possible that the same dynamic plays a role in the current model as well. Although the effect of selection on perspective-inference is initially strengthened by the egocentric bias because this bias exaggerates the fitness differential between variants under this selection pressure, it is possible that learners' egocentric bias further contributes to strengthening selection by transforming lexicons in the direction of becoming less informative (perhaps specifically towards lexicon types 2, 5 and 7, which act as attractors as shown in Appendix D), thereby eliminating variation from the population and making transmission more faithful. To explore whether this effect may be at play in addition to the effect of the egocentric bias on differential fitness, figures 4.9 and 4.8 show how the average lexicon variation (i.e. the number of variants in the population) and the average transmission fidelity change over the first 50 generations.

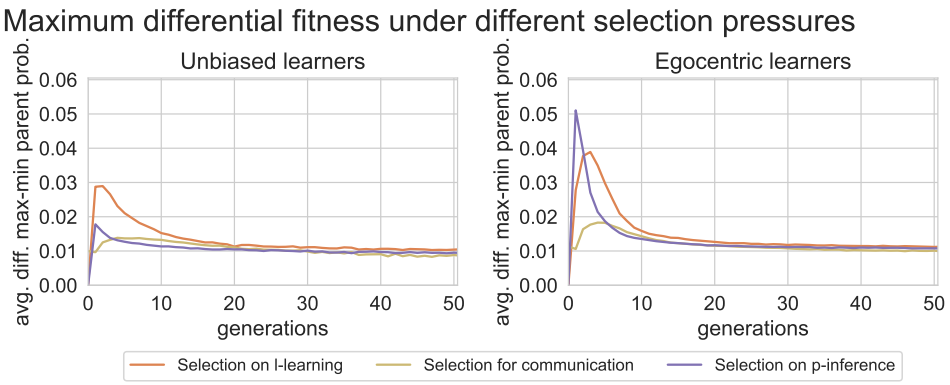


Figure 4.7: Average fitness variation in the population over generations, measured as the difference between the most fit and the least fit agent in each generation. Fitness is defined here as an agent's probability of being chosen as a cultural parent (i.e. to provide data for the next generation). Lines show mean over 100 independent simulation runs. The *No selection* condition is not depicted because fitness is undefined in the absence of a selection pressure (i.e. the probability of agents being chosen as a cultural parent is always uniform over the population in this condition).

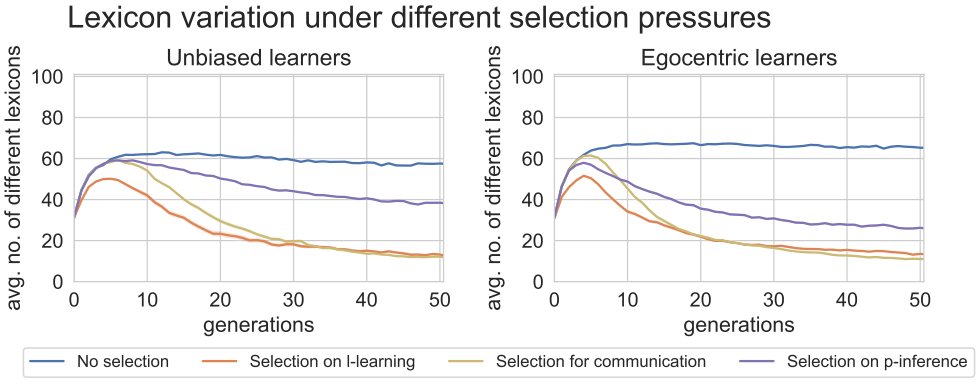


Figure 4.8: Average number of different lexicons present in the population over generations. Lines show mean over 100 independent simulation runs. Given that populations consist of 100 agents and that the total number of possible lexicons is 343, the maximum number of different lexicons in any given generation is 100. Grouping lexicons by lexicon type (i.e. informativeness category) shows yields the same pattern of results.

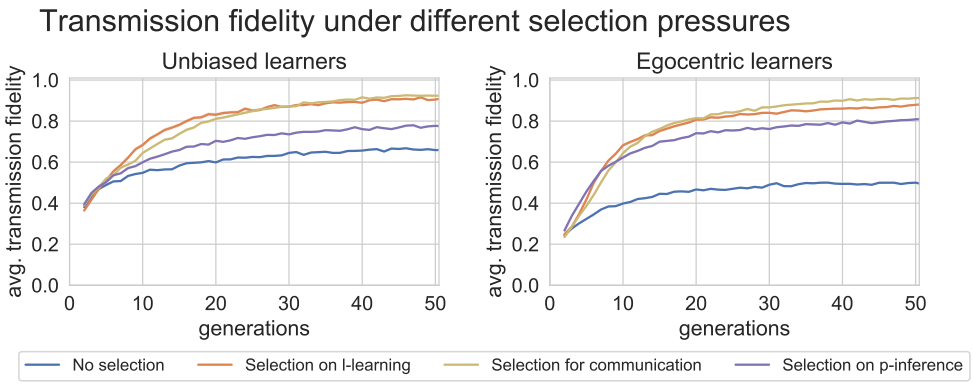


Figure 4.9: Average transmission fidelity for lexicons over generations. Lines show grand means over 100 independent simulation runs in a population of 100 agents (i.e. 10,000 transmission events per generation). No values are shown for first generation because this generation is initiated with a fixed set of lexicons which are not transmitted from any previous generation.

Figure 4.8 shows firstly that in the *No selection* condition, egocentric populations retain higher levels of lexicon variation than unbiased populations do. Thus, the convergent transformations in the direction of less informative lexicons caused by the egocentric bias does not lead to an elimination of variation in the population in the current model, but instead slightly increases variation. In the *Selection on lexicon-learning* and *Selection for communication* conditions, lexicon variation is reduced, and the extent and rate of this reduction follow the same pattern as the reduction in fitness variation shown in Figure 4.7. That is, lexicon variation is reduced somewhat more rapidly in egocentric populations compared to unbiased populations, and reaches a stable low level which is virtually indistinguishable between the two selection pressures and the two types of population (unbiased or egocentric). The *Selection on perspective-inference* condition shows a different pattern however: lexicon variation is reduced further in egocentric populations than it is in unbiased populations. This is in line with how the average informativeness develops in these selection*bias conditions as shown in Figure 4.3, and the distribution over lexicon types after convergence as shown in Figure 4.5. These figures showed that the average informativeness increases more in the *Selection on perspective-inference**egocentric condition than in the *Selection on perspective-inference**unbiased condition, by virtue of populations converging more exclusively on the most informative lexicon type in the former condition than in the latter. In addition, figures 4.3 and 4.5 show that the *Selection on perspective-inference**unbiased condition retains higher levels of informativeness variation than the

*Selection on perspective-inference**egocentric condition.[9]

Finally, Figure 4.9 shows how the average transmission fidelity develops over generations in the different selection*bias conditions. Firstly, we see that in the *No selection* condition, the average transmission fidelity is lower in egocentric populations than it is in unbiased populations. Despite this difference, transmission fidelity increases to nearly equal levels between unbiased and egocentric populations in the *Selection on lexicon-learning* and the *Selection for communication* conditions.

In the *Selection on perspective-inference* condition in contrast, transmission fidelity reaches a higher level in egocentric populations than it does in unbiased populations. This is at odds with the fact that egocentric learners are more likely to transform the lexicon they receive as input than unbiased learners. However, it is in line with the development of the lexicons in the population in these two selection*bias conditions as discussed above (and shown in figures 4.3 and 4.5). The *Selection on perspective-inference**egocentric condition leads to higher average informativeness and a stronger representation of the most informative lexicon type in the population, and, as shown in Figure 4.2, more informative lexicon types take (as a rough generalisation) fewer observations to be acquired correctly. Thus, the *Selection on perspective-inference**egocentric condition shows higher levels of transmission fidelity than the *Selection on perspective-inference**unbiased condition, because the interaction between learners' bias and the selection pressure leads to a stronger effect of selection in terms of increasing informativeness in the former condition than in the latter, and more informative lexicons lead to more faithful transmission. This 'side-effect' of the effect of selection being stronger in the *Selection on perspective-inference**egocentric condition than in the *Selection on perspective-inference**unbiased condition may cause a positive feedback loop: selection increases the amount of maximally informative lexicons in the population, thereby increasing transmission fidelity, which in turn may bolster the effect of selection further (because selection relies on low levels of transformations). This prediction is hard to test however because the effects of increased fitness variation and increased transmission fidelity are difficult to disentangle in the current model.

[9]The same pattern of results is found across conditions when lexicons are grouped by lexicon type (i.e. informativeness category).

4.4 Discussion

This chapter presented simulation results from an iterated learning version of the developmental model described in Chapter 3. The learning results in Chapter 3 showed that when Bayesian learners are faced with the joint inference task of having to simultaneously infer a speaker's lexicon and perspective from observations of the speaker's utterances in context, this leads to co-development of these two skills. That is, learners can accurately learn both attributes of the speaker given a sufficient amount of observations made in different contexts, but this learning is only successful if (i) the learner is able to represent the speaker's perspective and (ii) the speaker uses a lexicon that is at least somewhat informative (i.e. not completely ambiguous). In the current chapter, this model of development was embedded in an iterated learning model in which lexicons are passed on over generations, in order to explore under what circumstances a population of agents who develop in this way can evolve an informative lexicon from scratch. (That is, all simulations presented in this chapter started out with a population of agents who all share the same completely ambiguous lexicon, which associates every possible signal with every possible referent.)

In order to answer this question, this chapter compared a condition without any selection pressure (the *No selection* condition) to three selection conditions in which agents were selected as cultural parents (i.e. to provide data for the next generation) based on different criteria. Firstly, in the *Selection on lexicon-learning* condition, agents were selected based on whether they had correctly learned the lexicon of their own cultural parent. Secondly, in the *Selection for communication* condition, agents were selected on their success at interpreting the signals of their cultural parent during communication. Thirdly and finally, in the *Selection on perspective-inference* condition, agents were selected based on whether they had correctly inferred their cultural parent's perspective. In addition to these different selection conditions, the current chapter (mirroring Chapter 3) compared populations of unbiased learners with populations of egocentric learners. These egocentric learners start out with an unhelpful egocentric bias in favour of the hypothesis that their cultural parent will share their perspective. This bias is unhelpful because in all simulations reported in this chapter, cultural parents instead always had the opposite perspective to that of their learner.

The simulation results presented above show that when no selection pressure is present, populations do not converge on more informative lexicons than what would be

expected by chance (i.e. if agents are selecting lexicons at random). When subjected to selection on lexicon-learning in contrast, populations quickly converged on the most informative lexicon type (i.e. other types of lexicons were selected very rarely after convergence). However, as discussed in Section 4.3.1, this result is not very robust: when parameter settings change to create a more realistic situation, such as a larger lexicon or a tighter transmission bottleneck, the *least* informative lexicon type will come to dominate the population instead. This is because lexicons of the least informative type, which only contain completely ambiguous signal-referent mappings, do not require learners to correctly infer their parent's perspective in order to be learned accurately (as shown in Chapter 3). Thus, a pressure for accurately learning the lexicon of the previous generation does not robustly drive populations to evolve an informative lexicon, because accurate lexicon-learning does not necessarily require a learner to accumulate information about their cultural parent's referential intentions. When a lexicon maps every signal to every referent, it can be learned without knowing which particular referent a signal was intended to refer to when it is used; the learner only has to observe that all signals are used equally frequently regardless of the context.

When populations are subjected to selection on communicative success, they also quickly converge on the most informative lexicon type. This happens because this lexicon type, when shared between two agents, provides a hearer with the most unambiguous information about the speaker's intended referent, thereby leading to the highest possible communicative success. The latter will be the case regardless of any of the parameter settings (informativeness is in fact defined as the communicative success of a lexicon with itself), which means this result should be more robust to changes in parameter settings. Finally, when populations are subjected to selection on perspective-inference success, they also converge on the most informative lexicon type, although not quite as strongly as in the *Selection for communication* condition. In the *Selection on perspective-inference* condition, this outcome results from the fact that more informative lexicon types provide agents with more information about each others' perspectives. This is for the same underlying reason as the fact that more informative lexicon types lead to higher communicative success: a more informative lexicon provides more information about the speaker's intended referent. Inferring another agent's perspective is a matter of tracking how frequently the agent chooses to talk about the different referents in different contexts. Therefore, the more one-to-one signal-referent mappings a lexicon contains (and thus the more unambiguously the speaker's utterances are tied

to the speaker's intended referents), the easier it becomes to track the frequencies of the speaker's intended referents in different contexts. Just like for communicative success, this rule holds independently of the parameter settings. Therefore, convergence on the most informative lexicon type in the *Selection on perspective-inference* condition will be relatively robust to changes in parameter settings, just as in the *Selection for communication* condition. Under selection on perspective-inference, it is the cultural evolution of an informative lexicon type that leads to agents becoming more successful at inferring each others' perspectives, *not* any biological adaptation of agents' innate ability to learn about perspectives. Thus, this model provides a purely cultural evolutionary account of the evolution of improved perspective-inference alongside an evolving lexicon.

In sum, if the ability to communicate or the ability to infer others' perspectives is not relevant to a population, an informative lexicon does not evolve under the current model. However, if *either* communication or perspective-inference matters to how likely agents are to become a cultural parent, populations are able to establish an informative lexicon type from scratch. This happens despite the fact that individual learners' lexicon-learning and perspective-inference co-develop in this model. That is, populations are able to evolve an informative lexicon from scratch even if the way in which their lexicon and perspective-inference co-develops means that they need to receive input from an informative lexicon in the first place in order to accurately learn such a lexicon. The combination of selection and a transmission bottleneck solve this interdependency problem: the bottleneck (combined with a small amount of noise in transmission) means that in the first generation some agents will transform the lexicon they received input from (which is completely uninformative for every agent in the initial generation) in a way that makes the lexicon they infer somewhat more informative. Once such transformations have happened, differential 'fitness' (in the sense of agents' probability of becoming a cultural parent) is created in the subsequent generations because those agents who receive input from a more informative lexicon will be more successful at communicating or inferring perspectives (depending on the selection pressure at play). Once such differential fitness exists in the population, selection causes it to converge on the most informative lexicon type for as far as is necessary to eliminate lexicons that lead to lower levels of fitness from the population.

Under selection for communication, the emergence of a highly informative lexicon type causes populations not just to become very successful at communicating, but

also to become highly successful at inferring each others' perspectives. Similarly, under selection on perspective-inference, the emergence of a highly informative lexicon type causes populations not just to become highly successful at inferring each others' perspectives, but also to become relatively successful communicators (although not quite to the same extent as under selection for communication). Thus, we see that a co-evolutionary dynamic arises between lexicon-evolution and agents' perspective-inference success. This happens *by virtue* of the co-developmental model described in Chapter 3. Firstly, this model assumes that agents' perspective on the world always plays a role in communication, and that in order to learn a lexicon accurately, learners thus need to acquire knowledge about the speaker's perspective simultaneously with acquiring the lexicon. Secondly, the model assumes that these two attributes of the speaker have to be inferred based on the same data: the speaker's utterances in context. Learners can acquire information about the speaker's perspective only if the speaker's lexicon is at least somewhat informative; i.e. provides at least some information about the speaker's intended referents in the different contexts. However, as soon as a lexicon becomes somewhat informative, the speaker's utterances start depending on the interaction between the speaker's (unobservable) perspective and the (observable) context. As a consequence, a learner needs to acquire knowledge about the speaker's perspective in order to accurately learn any lexicon type that is not completely uninformative. Taken together, these two assumptions thus result in a model of communication and learning where the best solution to both successful communication and successful perspective-learning is to establish an informative lexicon. This in turn has as a result that when a selection pressure for improving one skill (be it communication or perspective-inference) leads to the population establishing an informative lexicon, the other skill that was not selected on will improve along with the first.

In sum, a co-evolutionary dynamic between the evolution of a lexicon and the evolution of better perspective-inference falls out naturally when the assumptions of the developmental model as described above are combined with iterated learning, as long as populations are under some pressure to evolve *either* successful communication *or* successful perspective-inference. In actual hominin evolution, either pressure may have come from a need for improved social coordination that came with foraging strategies changing to collaborative foraging and big game hunting during the Pleistocene, as discussed in Chapter 2 (Sterelny, 2012; Tomasello et al., 2012; Whiten and Erdal, 2012). However, the results presented in this chapter suggest that by virtue of the co-

evolutionary dynamic between lexicon-evolution and improved perspective-inference, these two pressures do not need to have both been present as a result of external factors. Because the further sophistication of one of these skills leads to further sophistication of the other, a pressure for evolving one can also cause an advance in the evolution of the other, in line with the two-way positive feedback loop between language and mindreading suggested by Sterelny (2012) and Whiten and Erdal (2012).

One could argue that initial selection on perspective-inference is more plausible than initial selection for communication, because for a useful communication system to have any pay-off, it needs to first be shared with other individuals. Selection on perspective-inference does not make this assumption of coordination between individuals. That is, being able to infer others' perspectives provides an individual with an advantage because it enables her to predict and manipulate others' behaviour, and this advantage holds regardless of whether other group members have the same mindreading abilities. The simulation results presented in this chapter suggest that such selection on perspective-inference could in turn drive selection for useful communication systems.

The fact that the model presented in this chapter is an iterated learning version of a Bayesian model that involves joint inference, leads to an unusual outcome compared to other iterated learning models with Bayesian agents. In the *No selection* condition, populations of egocentric learners do not converge to the learners' prior probability distribution over lexicons. This deviates from a general result that Griffiths and Kalish (2007) demonstrated for iterated learning chains of Bayesian agents that use sampling as their hypothesis selection strategy, which is that the stationary distribution over hypotheses converges to the learners' prior distribution.

When using the same method for numerically calculating the stationary distribution over lexicon hypotheses as described by Griffiths and Kalish (2007) (and sampling), the current model does lead to convergence to the prior in populations of unbiased learners (in line with the results of Griffiths and Kalish), but deviates from the prior in populations of egocentric learners. In egocentric populations, the less informative lexicon types are overrepresented in the stationary distribution, while the more informative lexicon types are underrepresented. This is a result of the fact that learning involves joint inference over a combination of two hypothesis spaces (lexicons and perspectives), and that learners can thus have a uniform prior over lexicons while at the same time having a nonuniform prior over the full space of composite hypotheses. Egocentric learners' prior over the hypothesis space of the variant that is culturally transmitted

through iterated learning (the lexicon) is uniform just like it is for unbiased learners. However, their prior over perspectives is strongly biased in favour of their own perspective. Perspectives are not culturally transmitted in this model, and are thus not taken into account in the stationary distribution over hypotheses. However, learners' prior over perspectives does influence the outcome of their learning, and thereby how likely they are to select different lexicons.

Egocentric learners' bias over perspectives causes them to have a tendency to underestimate the informativeness of the lexicon they are receiving input from. This is a result of an interaction between two factors. Firstly, egocentric learners require more observations to accumulate a given amount of posterior belief in the correct lexicon hypothesis than unbiased learners do. Secondly, within the space of lexicon hypotheses that learners consider in this model, the informativeness level of the least informative lexicon type (0.33... *ca*) is closer to the average informativeness of all lexicons together (0.5 *ca*), than the informativeness level of the most informative lexicon type (0.9 *ca*).Therefore, given the same amount of observations, an egocentric learner will on average underestimate the informativeness of their input lexicon more than an unbiased learner will. When this is combined with a bottleneck on transmission, it leads to egocentric learners being slightly more likely to select a less informative lexicon than unbiased learners, which is ultimately reflected in the stationary distribution over lexicons.

Navarro et al. (2018) showed that the relationship between a population's stationary distribution over hypotheses and the learners' prior as shown by Griffiths and Kalish (2007) is also distorted when populations are heterogeneous in terms of the strength of the agents' prior bias. As Navarro et al. (2018) discuss, such cases have important implications for how the iterated learning paradigm is used: it means that reconstructing learners' inductive bias based on the outcome of an iterated learning experiment is not straightforward and in certain cases not possible. The current model suggests that iterated learning with Bayesian agents who have to perform joint inference on two hypotheses of which only one is culturally transmitted, might present another such case. The current model was not designed to investigate the outcome of iterated Bayesian learning with joint inference on a general level, and the results found here may be quite specific to how lexicons are represented in the current model. However, iterated learning with joint inference in which one part of the composite hypothesis is culturally transmitted while the other is not, presents a case that warrants further research.

Finally, the *Selection on perspective-inference* condition presented in this chapter interacts with learners' egocentric bias in a different way than the other two selection conditions. Agents' fitness under selection on lexicon-learning or selection for communication depends only in part on whether they inferred their cultural parent's perspective correctly (in so far as this is relevant for correctly inferring the parent's lexicon or for correctly interpreting the parent's utterances, respectively). In contrast, agents' fitness in the *Selection on perspective-inference* condition depends solely on whether or not they correctly inferred their cultural parent's perspective. Thus, out of all three selection conditions, this is the condition in which agents' fitness is influenced most strongly by having an egocentric perspective bias. In the other two selection conditions, the unbiased populations converge more strongly on the most informative lexicon type than egocentric populations do, because lexicons are transmitted slightly more faithfully in unbiased populations than they are in egocentric populations, thus strengthening the effect of selection. (This is because, as discussed in Section 4.1.2, selection relies on high-fidelity transmission.) However, in the *Selection on perspective-inference* condition, this result flips: egocentric populations converge more strongly on the most informative lexicon type than unbiased populations do. This is due to the fact that for egocentric learners, receiving input from a highly informative lexicon is more important for reaching a high level of fitness than it is for unbiased learners. Or to put it in more general terms: evolving a useful language becomes more important when learning about others' minds is hard.

On a surface-level this result seems similar to that demonstrated by Claidière et al. (2018), who showed that when learners have a bias in the opposite direction of a selection pressure, the convergent transformations resulting from this bias can nevertheless strengthen the effect of selection compared to random transformations. Claidière et al. showed that in their experiment and model, this is due to convergent transformations causing higher-fidelity transmission by bringing variants closer to the attractor. Although in the current model we find a similar result: agents' egocentric perspective bias which leads to convergent transformations in the opposite direction (lower informativeness) than selection (higher informativeness) cause selection to have a stronger effect on the final distribution over lexicons compared to when learners are unbiased.

However, in the current model this effect is not due to the convergent transformations making transmission more high-fidelity, but more directly to the fact that learners' egocentric bias changes the fitness landscape such that the same difference in informa-

tiveness between two lexicon types will cause a bigger difference in agents' fitness in egocentric populations than it does in unbiased populations. To reiterate, evolving a useful lexicon becomes more important when learning about others' perspectives is hard. This direct effect of learners' egocentric bias may however cause a positive feedback loop between selection and transmission fidelity. Firstly, learners' egocentric bias makes it more important to their fitness to receive input from an informative lexicon, and therefore selection causes the amount of highly informative lexicons in the population to increase. Secondly, more informative lexicons facilitate their own high-fidelity transmission. Not because they fit the learners' biases better (although the learners' egocentric bias do make them more likely to evolve in this selection condition), but because more informative lexicons require fewer observations to be learned accurately (roughly speaking). Rather than being a result of learners' prior bias, the latter is a result of the likelihood. The more one-to-one mappings a lexicon contains, the less the data it produces will be confusable with that of other lexicon types.

In the model of communication used in this chapter, agents in the role of listener use their perspective-taking abilities to help them interpret a speaker's utterances. However, when the same agents take on the role of speaker, they do not take into account that other agents will be such perspective-taking listeners as well. This is not in line with the minimal requirements for ostensive-inferential communication as discussed in Chapter 2. In the next chapter, I will therefore present a model of communication which adds an existing and well-tested model of pragmatic reasoning on top of agents' perspective-taking abilities in order to turn them from literal speakers (who base their utterances purely on their own lexicon) into pragmatic speakers (who base their utterances on a model of the listener).

Chapter 5

Cultural evolution of lexicons in populations of pragmatic agents

In Chapters 3 and 4, speakers' utterance selection was implemented in the simplest possible way: given an intended referent, the speaker simply chose randomly between all signals that were associated with that referent in their lexicon. In other words, Chapters 3 and 4 used a literal speaker. A model of utterance interpretation (i.e. listening) was introduced in Chapter 4 for the purposes of the *Selection for communication* condition, because agents in this condition are selected based on how successfully they interpret the utterances of their cultural parent. Because these agents have just inferred their cultural parent's perspective through observational learning, the model of utterance interpretation used in Chapter 4 is that of a *perspective-taking listener*. In contrast to a literal listener, such a perspective-taking listener can resolve ambiguity in the lexicon by considering how likely the speaker is to choose each of the possible referents as their intended referent in the current context. (E.g. 'notebook' uttered by a speaker who's ready to sketch an idea is likely to refer to something different than 'notebook' uttered by a speaker who's wanting to check her email.) As will be discussed in more detail below, this ability to reason about the speaker's perspective gives perspective-taking listeners an advantage over literal listeners (provided that the listener's model of the speaker's perspective is accurate).

However, the fact that Chapters 3 and 4 used literal speakers means that the role of mindreading (in the sense of agents reasoning about each other's mental states) is limited. Agents in the role of listener make use of their ability to take other agents' perspective, but in the role of speaker they do not reason about other listeners being

such perspective-takers as well. The current chapter explores how the learning and cultural evolutionary dynamics change when speakers do have the ability to reason about other agents as perspective-takers. This is implemented by adding on an extra layer of pragmatic reasoning on top of both the speaker model and the listener model used in Chapters 3 and 4.

Below I will first review the existing models of pragmatic communication that this extension to the model is based on (Section 5.1), followed by a description of the model itself (Section 5.2). In Section 5.3 I will then present simulation results of exactly the same simulations of development and cultural evolution that were presented in Chapters 3 and 4, with the only difference that agents are now fully pragmatic speakers and listeners. Finally, in Section 5.4 I will discuss the findings of these simulations in relation to the findings reported in Chapters 3 and 4, as well as in relation to existing models of the evolution of linguistic conventions in populations of pragmatic agents and the broader aim of this thesis.

5.1 Review of models of pragmatic communication

Models of pragmatic communication attempt to describe and explain patterns in language use where meaning is conveyed that goes beyond the literal meaning of an utterance. Most of these models therefore study phenomena that involve interlocutors reasoning about each other's mental states. There are two main strands of models of pragmatic communication: game theoretic models, and probabilistic reasoning models. In this section I will first give a brief overview of each, followed by a discussion of their similarities and differences. I will then turn to models that have combined pragmatic communication with lexicon-learning and learning about other attributes of the speaker, followed by a discussion of an existing model of the co-evolution of pragmatic communication and linguistic conventions.

5.1.1 Game theoretic models of pragmatics

Game theory is well-suited for modelling pragmatic communication because it is a framework for describing interactive decision making, where one agent's decision may influence what the best next move is for the other agent. Within game theory, the *signaling game* (Lewis, 1969) describes coordination problems that can be solved by forming conventions. In this game, the speaker (normally referred to as the 'sender'

in game theory) knows the state of the world, but the listener ('receiver') does not. The speaker then has to select a signal to send to the listener, and if the listener interprets this signal by choosing the action that is appropriate for the current world state (as defined by a utility function), the communication event was successful. An *equilibrium* is a combination of sender and receiver strategies for which neither party could increase their pay-off by changing their own strategy. Game theory therefore offers a straightforward framework for analysing what optimal or rational communicative strategies look like. This is not limited to analysing what would constitute a useful set of linguistic conventions; it also makes it possible to analyse what rational communicative behaviour should look like given an already existing set of linguistic conventions (in the sense of literal semantic meaning). The latter is what game theoretic models of pragmatics usually focus on (see e.g. Franke, 2017, for a review), because they study reasoning and inference that goes beyond conventional semantic meaning. However, more recently these models have also been combined with cultural evolutionary models of convention-formation (Brochhagen et al., 2018).

In a recent review of game theoretic models of pragmatics, Franke (2017) distinguishes three different (but interlinked) approaches within this strand of modelling work. Firstly, the *evolutionary perspective* thinks of pragmatic phenomena as the result of gradual adaptation, habitualisation or conventionalisation through cultural evolution. Secondly, the *rationalistic perspective* focuses on what rational agents *should* do in order to solve a certain communication problem. And finally, the *probabilistic reasoning perspective* focuses on the speaker's and listener's reasoning processes and (in contrast to the rationalistic approach) allows for limitations and biases in that reasoning. I discuss this last perspective under the probabilistic reasoning strand of models of pragmatics in Section 5.1.2 below. Franke (2017) further argues that what unifies these different perspectives is the fact that they consider signal production and signal comprehension side-by-side, as attuned to each other in an interdependent system. As mentioned above, this interdependency of speaker and listener strategies is what motivates the use of game theory as a modelling framework.

Franke (2017) shows that each of these modelling approaches can help us understand the pragmatic implicatures that we see in actual language use, such as scalar implicature. Firstly, the *evolutionary perspective* can be implemented by applying the replicator dynamic model (briefly discussed in Chapter 4, Section 4.1.2) to a communication task by defining the fitness of an agent as the pay-off of the agent's strategy

relative to other strategies. The replicator dynamic then provides a simple formula for updating agents' strategies according to their fitness. Starting from a speaker and a listener with literal strategies for the task of successfully communicating the meanings SOME and ALL, Franke (2017) shows that, over a number of time steps, the replicator dynamic gives rise to a system in which the use of a signal that can mean both SOME and ALL is associated with the meaning 'SOME but not ALL'. That is, it recapitulates the pattern associated with the standard scalar implicature from SOME. The same approach yields similar results for other well-studied implicatures. Thus, this evolutionary perspective demonstrates how rational communicative behaviour can arise even if agents are themselves not rational, if we assume a selection pressure for communication.

Secondly, the *rational perspective* can be implemented within the game theoretic framework as the *iterated best response* model. Just like the evolutionary approach outlined above, this model starts from a literal production and comprehension strategy, but instead of these being the starting point for an evolutionary trajectory, they are simply a hypothesis that forms a starting point for pragmatic reasoning. That is, even if no agent actually uses a literal strategy, this strategy still forms the 'baseline' for pragmatic reasoning. A rational agent then maximises the expected utility of their strategy given their beliefs about what the interlocutor's strategy is. An action which maximises the agent's expected utility in this way is called a *best response*. The iterated best response model defines how rational agents should update their strategies as a result of a chain of iterated reasoning consisting of nested levels of beliefs about the interlocutor's strategy. Franke (2017) shows that this model gives rise to most of the same well-studied pragmatic implicature phenomena that the evolutionary approach does. The *iterated quantal response* model is a variant of the iterated best response model, where the best response is replaced with a probabilistic approximation of it. This type of model thus implements approximate rational choice with occasional errors, where the probability of making such errors is governed by an optimisation parameter. Franke shows that within a reasonable range of values for this optimisation parameter, the iterated quantal response model gives rise to all the standard implicatures that the evolutionary model does, including those that prove problematic for the iterated best response model. Thus, the iterated quantal response model demonstrates how the assumption that agents are approximately rational reasoners can give rise to the same implicatures as the evolutionary model does, without having to assume adaptation over evolutionary time.

Thirdly and finally, the *probabilistic reasoning perspective* departs not from a utility function but rather from the speaker's beliefs about the listener, followed by a chain of reasoning about nested beliefs similar to that of the *IxR* models (umbrella term for iterated best response and iterated quantal response models; Franke and Jäger, 2014) described above. Most models of this approach take the *rational speech act* model as a starting point, which I will discuss in more detail below, but they can equally be implemented in a game theoretic framework (see Franke and Jäger, 2016; Franke, 2017).

5.1.2 Probabilistic reasoning models of pragmatics

Probabilistic reasoning models of pragmatics capture communication in terms of speakers' and listeners' beliefs about each other, and aim to explain pragmatic communication using considerations of rationality or optimality (Franke and Jäger, 2016). The most extensive line of research of this type to date is within the framework of the *rational speech act* model (see Frank and Goodman, 2012 and Goodman and Stuhlmüller, 2013 for the original development of this model, and Goodman and Frank, 2016 and Frank et al., 2017 for recent reviews). In this brief overview of probabilistic reasoning models of pragmatics I will therefore focus on the rational speech act (henceforth RSA) model, but see Franke and Jäger (2016) for a discussion of this modelling approach in more general terms. The RSA model implements a speaker as an approximately rational agents who chooses their utterances to (soft)maximise their expected utility. The RSA model is often used to describe and predict pragmatic reasoning in reference games (Frank and Goodman, 2012; Frank et al., 2017), but has been used to model other phenomena in language use as well, as we will see below. In the context of a reference game however, the most basic formulation of the RSA model defines a speaker who chooses their signals proportional to how likely the listener is to infer the intended referent given the signal (Frank et al., 2017). A speaker of pragmatic reasoning level n does this by using Bayesian inference to 'invert' the interpretation procedure of a listener of level *n-1* (i.e. one level below the speaker in terms of pragmatic reasoning), to yield a set of corresponding production probabilities.

In most reference game versions of the RSA model however, the speaker's utility is defined not in terms of the listener's 'action' (e.g. whether or not the listener interprets the referent that the speaker intended) but rather in terms of the listener's

belief about the intended referent; specifically, how much information the listener gains about the intended referent when observing the speaker's utterance. This information-gain is defined as the negative surprisal of the intended referent given the signal for the listener; an information-theoretic measure of how much more certain the listener becomes about the intended referent after hearing the speaker's utterance.[1] In other words, the speaker's goal is to provide 'epistemic help' to the listener (Goodman and Frank, 2016). The fact that the speaker's goal is defined with respect to the listener's belief in the RSA model makes it different from models of pragmatics in the game theoretic tradition as described above, which define the speaker's utility in terms of the listener's action. That is, given a cooperative model of communication, a speaker in the game theoretic tradition is successful if the listener's action is appropriate for the current world state, and unsuccessful if not.

The level-n pragmatic listener in the RSA model reasons about a speaker of level *n-1* and, similarly to the pragmatic speaker but in the other direction, uses Bayesian inference to recover the probabilities of possible referential intentions given the speaker's utterance. The pragmatic listener reconstructs how likely each of the possible referential intentions is to have led the speaker to choose the utterance that she chose and not another one. Note that both pragmatic speakers and pragmatic listeners can be of any level n, and always reason about another agent one level of pragmatic reasoning below them (*n-1*). This recursion is potentially infinite, but the reasoning chain in the RSA model is normally grounded in semantic meaning by starting from a literal listener. A level-1 pragmatic speaker thus reasons about a level-0 *literal* listener, who interprets utterances by simply choosing between each of the interpretations that is compatible with the utterance with uniform probability. The interpretation probabilities of this literal listener are thus directly and solely determined by the lexicon. Because the pragmatic listener interprets utterances by reconstructing how likely the speaker is to have had different possible communicative intentions given that she chose to use that utterance, it allows for simulating behaviour in situations where the literal meaning of the utterance (as defined by the lexicon) leaves ambiguity about the underlying communicative intention.

[1]The negative surprisal of the intended referent r given signal s for a listener of level *n-1* is defined as $ln(P(L_{n-1}(r|s)))$. A speaker who softmaximises this negative surprisal (defined as $P_{S_n}(s|r) \propto e^{\alpha ln(P(L_{n-1}(r|s)))}$ minimises the Kullback-Leibler divergence between their own belief in the intended referent (which is simply $P(r_{intended}) = 1.0$) and the listener's belief in the intended referent (Franke, 2017). (Kullback-Leibler divergence is a standard information-theoretic measure of the difference between two probability distributions.)

Frank and Goodman (2012) first introduced the RSA model to predict people's communication behaviour in a simple referential communication game in which speakers had to choose a single word to pick out an object from a context where several objects could align on several feature dimensions. For instance, a speaker might have to choose between *blue* and *circle* to pick out a blue circle in the context of a blue square and a green square. Or a listener might receive the word *blue* in this context and have to infer the most likely referent. The version of the RSA model designed to simulate pragmatic behaviour in this task defined the speaker's utility as the amount of information gained about the referent by the listener (i.e. negative surprisal) as described above. In the case of this task, informativeness is equal to the specificity of the signal. That is, the most informative strategy is to describe a feature of the intended referent that it shares with as few other objects as possible. Frank and Goodman (2012) showed that the predictions of this RSA model correlate highly with the actual choices of human speakers and listeners in this task (when combined with an empirical measure of the a priori salience of the three different referents[2]).

Goodman and Stuhlmüller (2013) extended the RSA model to simulate scalar implicatures (the phenomenon that for instance "some of the apples are red" implies "some *but not all* of the apples are red" even though the meaning *all* is in principle compatible with the literal semantics of 'some'). Goodman and Stuhlmüller showed firstly that the RSA model can give rise to scalar implicatures, but also that these can be partially or fully cancelled if the listener knows that the speaker only has partial knowledge (i.e. if the speaker cannot see all the apples). A reduction in the speaker's information access affects their signalling behaviour by reducing their certainty about the state of the world. Importantly, the speaker's information access is assumed to be observable to the listener. The model predictions show that a pragmatic listener will have a very low belief in the speaker intending *all N* when she said 'some N', but that this belief increases as the speaker's information access decreases, down to *all N* being almost as likely as *some N* if the speaker could only observe one out of three objects. In an experimental version of this task, Goodman and Stuhlmüller asked participants to place a bet by dividing a fixed amount of money over possible outcomes (i.e. "How many of the 3 letters do you think have checks inside?") given an utterance and full knowledge

[2]However, Frank et al. (2017) discuss the finding that such empirical estimates of the prior probability of referents does not actually improve model fits of the RSA model to empirical data of reference games.

of the speaker's information access. The experimental results showed the same qualitative pattern as the pragmatic listener in the RSA model, and a tight quantitative fit of the model could be produced by parameterising the assumed degree of optimality of the speaker and the listener's prior belief in all objects having the property of interest (i.e. the base rate of all apples being red). This study thus demonstrates that the RSA model can incorporate a form of perspective-taking in pragmatic reasoning: that is, the listener's knowledge about the speaker's knowledge *access* (e.g. how many apples the speaker can see) alters the listener's interpretation behaviour in a rational way.

5.1.3 Similarities and differences between game theoretic models and the rational speech act model

As discussed by Franke and Jäger (2014), there are strong similarities between the game theoretic IxR models and the RSA model. The way in which speakers and listeners reason about each other's production and reception behaviour in a chain of nested beliefs is the same in both types of model. However, there are also some conceptual differences between these two strands of modelling (Franke and Jäger, 2014; Franke, 2017). Firstly, as mentioned above, agents in the IxR models are concerned first and foremost with each other's *actions*, while agents in the RSA model care only about each other's *beliefs*. That is, in the IxR model, the goal of the speaker is to induce the appropriate action in the listener for the current world state, as defined by a utility function. In the RSA model on the other hand, the utility of the speaker's possible utterances is always defined in relation to the listener's belief, for instance the listener's belief about the speaker's referential intention. Secondly (and relatedly), there is a difference in how listeners select their interpretations in the two types of models. The listener in the IxR model chooses an action (approximately) optimally to maximise their expected utility. The listener in the RSA model instead chooses each of the possible interpretations with probability proportional to their posterior belief in the hypothesis that that interpretation corresponds to the speaker's communicative intention. Thus, unlike in the IxR models, the listener in the RSA model does not have the direct goal to maximise their own expected utility, but rather to correctly infer the speaker's communicative intention. Thirdly, while IxR models consider both a literal listener and a literal speaker as possible starting points of the chain of pragmatic reasoning, RSA models usually only start from a literal listener (Goodman and Frank, 2016; Frank

et al., 2017).

5.1.4 Models of learning about lexicon and speaker in pragmatic agents

As summarised in Section 5.1.2, the RSA model provides a good fit for empirical data of pragmatic communication if the speaker and listener already share a lexicon. However, this thesis is concerned with the *learning* (individual and iterated) of attributes of the speaker, such as her lexicon. Below I will review variants of the RSA model that were developed to model lexicon-learning and learning about other attributes of the speaker in pragmatic agents, in turn.

Lexicon-learning in pragmatic agents

Frank et al. (2009a,b), and Frank and Goodman (2014) applied the RSA model to the task of word learning. Each of these models is based on the intuition that if a learner assumes that the speaker picks her utterances to be maximally informative, this can help the learner determine the meaning of a novel word. Frank et al. (2009a) and Frank and Goodman (2014) assume that the learner always has full knowledge of the speaker's intended referent, and only has to infer which *feature* of that referent the speaker's utterance is referring to, given the context. In other words, the learner knows the intended referent but has to infer the lexicon. For example, a learner who observes a speaker using a novel word to refer to a red circle in the context of a blue circle, will infer that this novel word means *red* rather than *circular*, because otherwise the utterance would be uninformative. Frank et al. (2009b) extend this model to one in which learners have to infer not only the lexicon, but also the speaker's referential intentions. Thus, learners in the Frank et al. (2009b) model face a joint inference task similar to that faced by learners in the model presented in this thesis.

Frank et al. (2009a) tested the predictions of the lexicon-inference-only model using two experiments designed to test the model's comprehension and production predictions respectively. In the comprehension experiment, adult participants were presented with novel words and told that these were uttered by a speaker of a foreign language who was trying to teach them the meaning of the words. Each word was presented together with a context of six simple objects which varied along two binary-valued dimensions[3],

[3]The two dimensions used in a given context were picked from a larger set of four possible dimensions, and contexts were constructed in such a way that participants could not use mutual exclusivity across trials to figure out the meaning of a novel word.

and participants were shown explicitly which of these objects the novel word was used to refer to. (E.g. a possible context is one red circle, one blue circle, and four blue squares.) As in the experiment of Goodman and Stuhlmüller (2013), participants were then asked to split a bet of "$100" (used to simulate degrees of belief) between two possible features that the novel word could refer to (e.g. *red* and *circular*). Frank et al. (2009a) found that participants' bets were highly correlated with the probabilities predicted by the model (in which the listener reasons about a rational speaker who maximises the informativeness of their signal given the context, as described above for the Frank and Goodman, 2012 model).

In Frank et al.'s (2009a) production experiment, adult participants were first familiarised with a large set of stimuli that could differ along many different dimensions. For this purpose, Frank et al. used pictures of 304 different collectible bouncing balls called 'superballs' for which an existing set of tag words describing their distinguishing features was available. Participants were then asked to write a short description in English for each ball out of a randomly chosen subset of 50 balls, "so that someone could pick it out of the full set". Frank et al. (2009a) found that except for basic-level colour terms, the words that participants used to describe these superballs were well predicted by the model. To recap, the model assumes a rational speaker who maximises the informativeness of their signals by choosing the most specific word to pick out the intended referent given the context. In sum, Frank et al.'s (2009a) experiments show that adult human speakers choose utterances in a way similar to the RSA model's predictions when instructed to be informative (i.e. to make sure that a listener would be able to pick out the referent given the context), and that adult human listeners interpret utterances in line with the RSA model's predictions when told that the speaker has the intention to teach them the meaning of the novel word.

In a different set of experiments, Frank and Goodman (2014) tested the word learning predictions of this same model not just on adults but also on pre-school children. As in the comprehension experiment of Frank et al. (2009a), participants in this study were presented with a novel word together with a context of several objects which varied along two binary-valued dimensions, and were shown explicitly which of these objects the novel word was used to refer to. Participants were then asked to guess which of the two features of the target object was the meaning of the novel word. In order to test not only the qualitative but also the quantitative predictions of the model, the adult group was presented with different conditions which manipulated the extent to which

the two features were shared between the different objects in the context. The adult group was asked to place bets on the two possible meanings of the novel word just as in Frank et al. (2009a), while the child group was given a forced-choice version of the task. Frank and Goodman (2014) found firstly that the responses of pre-school children in a simple single-condition version of this experiment followed the qualitative predictions of the model, and furthermore that the bets of adult participants in the multi-condition version correlated highly with the quantitative predictions of the model.

The model of Frank et al. (2009b), in which learners have to infer not only the lexicon but also the speaker's referential intentions, was described in more detail in Chapter 3 (Section 3.3.3). To recap, the learner in this model assumes that speakers have a referential intention which is a function of the physical context (although 'empty' intentions are also considered a possibility), and that this referential intention mediates between the context and the speaker's utterance. Thus, a speaker's utterance in a given context is determined in part by her referential intention and in part by her lexicon. Both the speaker's lexicon and the speaker's referential intention are unobservable to the learner however. Similarly to the model presented in this thesis, the learner only gets to observe the speaker's utterances in context. The learner then uses Bayesian inference to simultaneously infer the speaker's lexicon and referential intention (where the former is constant but the latter changes with the context). The hypothesis space of referential intentions that the learner considers in a given context consists of each possible subset of the objects present in the context, and an 'empty' intention (used to account for cases where the speaker uses a word to refer to something that is not physically present). Frank et al. (2009b) tested this model on learning a lexicon from annotated data from the CHILDES corpus (MacWhinney, 2000), and showed that it has higher precision than a set of comparison models.

In sum, the models of Frank et al. (2009a), Frank and Goodman (2014) and Frank et al. (2009b) show how the assumption that a speaker tries to be informative can help listeners infer the meaning of a novel word. That is, when the listener has uncertainty about the lexicon, their ability to model the speaker accurately helps them infer the lexicon. This relates to the findings of Parish-Morris et al. (2007) (discussed in Chapter 3, Section 3.1.2) showing that children use their ability to infer communicative intentions in word learning, even in the absence of other social cues such as eye gaze. The RSA model can thus straightforwardly model the role that mindreading plays in word learning.

Learning about other attributes of the speaker in pragmatic agents

Kao et al. (2014b,a) and Kao and Goodman (2015) have explored the consequences of the listener being uncertain about the speaker, rather than the lexicon. This subclass of rational speech act models was later dubbed the *uncertain rational speech act* (uRSA) model by Goodman and Frank (2016). In the uncertain rational speech act model, the listener considers several different utility functions for the speaker; i.e. different properties of the signal that the speaker could be maximising.

By incorporating uncertainty about the speaker's communicative goal, Kao et al. (2014b) were able to model a rational listener that can understand non-literal language use such as hyperbole (e.g. "The electric kettle cost a thousand dollars"). To model such hyperbolic statements, Kao et al. (2014b) added an extra dimension to the space of possible communicative intentions and allowed the speaker to maximise on *either* of these two dimensions (but not the other). More specifically, the speaker could either maximise on the factual dimension (i.e. the state of the world) or the 'affective' dimension (i.e. the speaker's feelings or opinion about the state of the world). The listener's task was then to infer not just what the speaker intended to communicate about the state of the world or their own state, but also which of these two dimensions she intended to communicate. In order to perform this joint inference task, the rational listener needs a probability distribution over possible states of the world, and a probability distribution over the possible affects an agent can have *given* the state of the world (e.g. how likely it is that a speaker would find a kettle expensive, given its price). These probability distributions simulate the common ground of speaker and listener, and Kao et al. (2014b) estimated each of these two distributions empirically using judgements made by human participants. Kao et al. (2014b) ran an additional two experiments to compare the model's predictions with human judgements about both hyperbole (i.e. exaggerated statements to communicate affect) and the so-called 'pragmatic halo' effect: the fact that people tend to interpret round numbers such as 1,000 approximately but sharp numbers such as 1,001 precisely. Kao et al. (2014b) found that the model's predictions showed a tight fit with human judgements on both types of non-literal language use.

Kao and Goodman (2015) extended this model in order to simulate understanding of irony, by further dividing affect into two separate dimensions: valence and arousal. In this model, the communicative goal of the speaker can be either to inform the listener

about the state of the world or about either of the two dimensions of their affect. This extended model is able to interpret an utterance like "the weather is amazing" as *the weather is terrible* if the listener assigns a high prior probability to the weather being terrible (based on observation of the state of the world). This is because by separating affect out into a valence dimension (negative vs. positive) and an arousal dimension (high vs. low), the pragmatic listener can infer that the speaker intends to communicate her level of arousal (high in this case) rather than the valence of her affect (negative in this case). Kao and Goodman (2015) estimated two prior probability distributions empirically using experiments: (i) the probability of the speaker holding a certain belief about the weather (e.g. 'terrible', 'bad', 'neutral' etc.) given a particular state of the world; and (ii) the probability of the speaker having a certain affect (separated into arousal and valence) about the weather given the state of the world (where world states consisted of pictures of different weather types). In a separate experiment, participants were then asked to rate different utterances on how likely they were to be intended ironically, given different weather pictures. Kao and Goodman (2015) fitted two free parameters of the model to the data from this last experiment: the speaker's optimality and the prior probability of the different communicative goals (i.e. (i) real state of the world, (ii) affect: valence, and (iii) affect: arousal). Given the best fit of these parameters to the data of the final experiment (and the empirically estimated prior probability distributions of the other two experiments) the model predictions correlated highly with participants' irony ratings.

Finally, Kao et al. (2014a) used a similar model to simulate interpretation of metaphors of the type "X is a Y" (e.g. "My lawyer is a shark") where the speaker's communicative goal is to make the listener associate only a subset of the features of Y with X (e.g. *scary* but not *finned*). In this model, the task of the listener is to judge to what extent the speaker intends to communicate each of three different features that are associated with the animal in the predicate (e.g. *scary*, *dangerous* and *mean* for "is a shark"). In order to make this judgement, the pragmatic listener combines three different prior probability distributions to come to an interpretation; (i) the probability that the subject of conversation belongs to a certain category (animal or human); (ii) the probability that particular features (e.g. *scary*, *dangerous*) would apply to a subject of that category; and (iii) the probability of the speaker having a specific communicative goal. The latter distribution was assumed to be uniform if the utterance was an answer to a vague question like "What is he like?" but not if it was an answer

to a specific question like "Is he scary?". The second distribution (the probability of the member of a given category having a certain feature) was estimated empirically. The first and third distributions were treated as free parameters that were fitted to experimental data of human participants judging whether certain features (e.g. *scary*, *dangerous*) applied to the subject of conversation based on an utterance like "He is a shark". Kao et al. (2014a) found that the predictions of the uRSA model fitted to the experimental data correlating significantly with the participants' data.

In sum, the (u)RSA model can make accurate (qualitative and quantitative) predictions of human communication behaviour both when the listener is uncertain about the lexicon (Frank et al., 2009a; Frank and Goodman, 2014; Frank et al., 2009b) and when the listener is uncertain about the speaker (e.g. the speaker's communicative intention in terms of what aspect of the utterance she wants to maximise) (Kao et al., 2014b,a; Kao and Goodman, 2015). However, this thesis is concerned with what happens when both types of uncertainty are present simultaneously. That is, when a learner has to acquire both knowledge of the lexicon and knowledge about how other the speaker works, when both these types of knowledge can inform each other.

Bello (2012) identifies three conditions that an accurate computational model of mindreading should satisfy; according to his analysis, a mindreading agent should have the ability to (i) ascribe mental states to other agents, (ii) predict other agents' behaviour on the basis of these ascriptions, and (iii) explain the behaviour of other agents using *post hoc* ascriptions. These three conditions return in models of pragmatic communication in (at least) the following ways: pragmatic agents have the ability to (i) ascribe a communicative intention to a speaker, (ii) predict how a listener will respond upon receiving a particular signal, and (iii) explain the linguistic behaviour of another agent by ascribing a particular lexicon and/or communicative intention to them. Conditions (i) and (ii) are satisfied in models of pragmatic communication in general, and condition (iii) is additionally satisfied in models of pragmatic lexicon-learning and the uncertain rational speech act (uRSA) model.

5.1.5 Co-evolution of lexicon and pragmatic ability

Brochhagen et al. (2018) present a model of the co-evolution between agents' pragmatic reasoning level and their lexicons. This model takes scalar implicatures (and specifically the inference from *some* to *some but not all*) as its test case, and starts from the question

how the division of labour between semantics and pragmatics that we observe in this test case could come about. The conventional analysis of scalar implicatures is that scalar items like *some* are underspecified in their semantic meaning (in the case of *some* this is the lack of an upper bound specifying that *some* is not compatible with *all*), and that the enrichment to *some but not all* arises instead from pragmatics. For instance, returning to our red apples, a listener who knows that the speaker has observed all the apples would infer that an informative speaker who knew that all the apples were red would have said so, because that would have been a more informative utterance; therefore, if the speaker used 'some', it must mean *some but not all*.

Brochhagen et al. specify agents as having two attributes: a lexicon and a level of pragmatic reasoning (either literal or level-1). Similarly to the model used in this thesis, Brochhagen et al. model agents as Bayesian learners whose hypothesis space consists of all possible combinations of lexicon hypothesis and 'communication type' (i.e. literal or pragmatic) hypothesis. Again similar to the model used here, the hypothesis space of lexicons comprises all logically possible lexicons consisting of binary mappings between three meanings (*none*, *some* and *all*) and three signals. The lexicons of interest for the question of Brochhagen et al. are (i) those that use one-to-one mappings between meanings and signals (such that there are two different signals for *some* and *all*, and the signal for *some* does not map to *all*), and (ii) lexicons that have separate signals for *none* and *all*, and a third signal that maps to both *some* and *all*. The first type of lexicon specifies an upper bound for *some* in its semantics (Brochhagen et al. therefore call it L_{bound}), while the second type does not (thus called L_{lack}).

As mentioned above, agents can be either of two types of communicators: literal or pragmatic. Literal speakers and listeners choose signals and interpretations purely based on their truth-value (i.e. whether or not they map to a corresponding meaning or signal respectively). The pragmatic speaker in contrast softmaximises the probability that a literal listener would interpret their signal as the intended meaning. This means that the speaker's utility in this model is defined in terms of the listener's *action* rather than the listener's belief, following the game-theoretic tradition of pragmatics models discussed above. Brochhagen et al. made this design choice partly on the basis that this model is concerned with the evolution of communication types and lexicons, and includes a selection condition in which agents' fitness depends on their communicative success, which can be argued to be more plausible if agents' communication is concerned with actions rather than beliefs. The other reason for this design choice was of a more

practical nature: Brochhagen et al. (2018) use Griffiths and Kalish's (2007) approach for numerically calculating the outcome of an iterated learning chain of Bayesian agents, as described in Chapter 4 (Section 4.3.1) which requires that there is a nonzero probability of each possible hypothesis (combination of lexicon and pragmatic level in this case) transitioning into each other possible hypothesis from one generation to the next. For this to be the case, it is necessary that both types of speakers occasionally make an error in production. In the case of Brochhagen et al.'s model, this will happen if the speaker's production probabilities are directly proportional to the posterior probability of the signal given the intended referent for the listener, but not if the pragmatic speaker instead uses negative surprisal as their utility function, as in the RSA model.

Brochhagen et al. (2018) find that the division of labour between semantics and pragmatics that we observe in the case of scalar implicatures in natural language (in this model, this division of labour corresponds to pragmatic agents with lexicon type L_{lack}) arises when at least two conditions are satisfied simultaneously. Firstly, there needs to be a pressure for communicative success (favouring L_{bound} over L_{lack}), which in the case of this model results from selection for successful communication. Secondly, there needs to be a pressure for learnability, which in this model is implemented as learners having an inductive bias for simplicity (favouring L_{lack} over L_{bound}). When combined, the trade-off between these two competing pressures can lead to populations converging on a single target type of pragmatic agents who have all converged on the same lexicon of type L_{lack}. However, the extent of this convergence depends on the optimality parameter which determines how optimal pragmatic speakers are, and a second parameter which determines the extent to which learners' hypothesis selection is biased in the direction of choosing the hypothesis with maximum a posteriori (MAP) probability (Brochhagen et al. use a parameter which interpolates between pure sampling and pure MAP, just like Kirby et al. (2007) as described in Chapter 4, Section 4.1.1). The more optimal the pragmatic speakers are (increasing their communicative success given L_{lack}), and the more likely agents are to select the lexicon with highest posterior probability to (causing an amplification of learners' inductive bias, as discussed in Chapter 4, Section 4.1.1), the more populations converge on only pragmatic communication and a single lexicon of type L_{lack}.

The model of Brochhagen et al. (2018) comes very close to what we want to model here: learners use Bayesian inference to infer two attributes of the speaker which jointly determine the speaker's utterances given her intended meaning. And these lexicons and

pragmatic levels are transmitted culturally over generations using iterated learning, under a selection pressure for communicative success. Although this is not explicitly explored in Brochhagen et al.'s paper, agents' lexicon-learning and learning about their cultural parents' communication type presumably go hand-in-hand: as information is gained about the parent's communication type, the lexicon should be easier to learn and vice versa. However, a mixture of communication types in a population in this model is in a sense taken to be a transition period: the 'target type' is a pragmatic agent with an L_{lack} lexicon type, and Brochhagen et al. are interested in the circumstances under which this target type will completely take over the population. In contrast, this thesis is concerned with how lexicon-learning co-develops and co-evolves with learning something about other agents which is (conceptually) a stable source of variation in the population: agents can have different perspectives and these are not culturally transmitted. If perspectives are a stable source of variation in the population, and it is useful for agents to be able to make inferences about these, it makes sense in the current model to track how learning about lexicons and perspectives co-develops and co-evolves, and to explore not just a selection pressure for communication, but also selection on perspective-inference.

5.2 An integrated model: combining perspective-taking with pragmatic reasoning

In Chapters 3 and 4, speakers produced utterances literally. A literal speaker chooses randomly between all signals that are associated with their intended referent according to their lexicon, with some small probability ($P(\epsilon) = 0.05$) of making an error, defined as choosing a signal that is not associated with the intended referent (see Equation 3.2 in Chapter 3), reproduced below as Equation 5.1).

$$P_{S_0}(s \mid r, \ell) = \begin{cases} \dfrac{1-\epsilon}{|s_r|} \text{ if } s \text{ maps to } r \text{ in } \ell \\ \dfrac{\epsilon}{|\mathcal{S}| - |s_r|} \text{ otherwise} \end{cases} \tag{5.1}$$

where $|s_r|$ denotes the number of signals that map to referent r in lexicon ℓ, and $|\mathcal{S}|$ denotes the total number of signals in ℓ.

Learners have an accurate model of how the speaker produces utterances, but have to infer the speaker's lexicon and perspective. When agents' communicative success

(after learning) is relevant for their fitness, as is the case in the *Selection for communication* condition in Chapter 4, listeners make use of their knowledge of the speaker's perspective during comprehension. Such a *perspective-taking listener* does this using the same Bayesian procedure as the pragmatic listener in the rational speech act (RSA) model. That is, in order to determine the probability that the intended referent is r given that the speaker used signal s, the perspective-taking listener uses their model of how the speaker chooses referential intentions and signals, and 'inverts' this model of the speaker using Bayes' rule, as shown in Equation 5.2 (ee also Equation 4.1 in Chapter 4).

$$P_{L_n}(r \mid s, \ell_{L_n}, c, p') \propto P_{S_{n-1}}(s \mid r, \ell_{L_n})\, P(r \mid c, p') \tag{5.2}$$

where L_n stands for a pragmatic listener with level-n reasoning above a literal listener (a perspective-taking listener is L_1), and S_{n-1} denotes a speaker of one level of pragmatic reasoning below L_n (in the case of the perspective-taking listener this is S_0; a literal speaker). The probability that such a literal speaker (S_0) will produce signal s given referent r and lexicon ℓ ($P_{S_0}(s \mid r, \ell)$) is shown in Equation 5.1 above (originally defined in Chapter 3 as Equation 3.2). Note however that listener L_n uses not the lexicon of the speaker, but instead their own lexicon ℓ_{L_n} when calculating the speaker's production probabilities. Thus, listener L_n assumes that the speaker shares their lexicon. This is not an odd assumption to make in the condition in which this perspective-taking listener is used: the *Selection for communication* condition. As described in Chapter 4 (Section 4.2.3) agents in this condition are evaluated on their success at interpreting the utterances of their own cultural parent, from whom they have just learned their lexicon using Bayesian inference with a uniform prior over lexicons. After observing data from this cultural parent, agents select their own lexicon by sampling from their posterior probability distribution. The perspective-taking listener's lexicon ℓ_{L_n} is thus not their *best* guess at what their cultural parent's lexicon is (which would be the case if the listener used maximum a posteriori hypothesis selection), but it is an informed guess.

In the case of reference game variants of the RSA model (such as Frank and Goodman, 2012), the prior probability of the referent $P(r)$ refers to some baseline saliency of the referent. As discussed in Section 5.1.2, Frank and Goodman (2012) measured this empirically, and subsequently used the resulting estimated saliencies for making model

predictions that could be tested against empirical data of production and comprehension behaviour in a reference game. In the current model however, the probability with which the speaker will choose a given referent r as their referential intention depends on both the context and the speaker's perspective. Therefore, the perspective-taking listener defined in Equation 5.2 uses $P(r \mid c, p')$ to determine the prior probability of the referent. Just as for the lexicon, L_n uses their own model of the speaker's perspective p', rather than the speaker's real perspective p, when calculating their interpretation probabilities.[4]

In sum, the perspective-taking listener used in the *Selection for communication* condition in Chapter 4 is similar to a level-1 pragmatic listener (L_1) in the RSA model, except that here the perspective-taking listener reasons about a literal speaker (S_0) instead of a level-1 pragmatic speaker who in turn reasons about a literal listener. In other words, whereas the RSA model normally bottoms out at a literal listener, the current model bottoms out at a literal speaker. The rationale behind this model decision is explained below. Furthermore, the perspective-taking listener in the current model uses their own lexicon and their model of the speaker's perspective in interpretation, both of which may not correspond to the reality about the speaker.

Thus, in the model presented in Chapters 3 and 4, listeners use their ability to take the speaker's perspective, but speakers in turn do not produce their utterances with such a perspective-taking listener in mind. This gives rise to a somewhat odd asymmetry, given that the role of speaker and the role of listener are implemented within the same agent. That is, agents use their perspective-taking ability when they take on the role of listener, but are oblivious to the fact that other agents are perspective-takers as well when they take on the role of speaker (or at least do not make use of the fact that the listener will take their perspective when they decide what signal to use in a given context). To resolve this asymmetry, the current chapter turns agents into level-1 pragmatic speakers (who reason about a perspective-taking listener) and level-2 pragmatic listeners (who reason about a level-1 pragmatic speaker), and explores how this changes agents' development and cultural evolution, using the same simulations as used in Chapters 3 and 4.

[4]The listener's model of the speaker's perspective p' is not equal to the listener's own perspective p_{L_n} because unlike lexicons, perspectives are not transmitted culturally.

5.2.1 Pragmatic communication

Level-1 pragmatic speakers and level-2 pragmatic listeners are implemented following the RSA model (Goodman and Frank, 2016). However, as mentioned above, the model of pragmatic reasoning used here bottoms out at a literal speaker, rather than a literal listener as is usually the case in RSA models. The current model of pragmatic reasoning starts from a literal speaker for two reasons. Firstly, because the model of learning used in this thesis is Bayesian inference, we want learners to have an accurate model of how a speaker with a given perspective and lexicon chooses referents and utterances, in order to accurately calculate the likelihood of the data under a given composite hypothesis about the speaker. Therefore, learners always need to do one level of pragmatic reasoning above that of the speaker they're receiving input from, otherwise they would not have an accurate model of how the data they're observing was generated. It is then only fair to allow these learners to subsequently use this same level of pragmatic reasoning during communication. Secondly and relatedly, starting from a literal speaker allowed us to start from the simplest possible model of a speaker whose communication behaviour is influenced by their perspective: a speaker whose referent choice is affected by their perspective, but whose utterance choice is literal. In the current chapter we can build on this baseline model and see how agents' development and the cultural evolution of lexicons changes when agents' reasoning about each other becomes more sophisticated. Either or both of these assumptions may be unwarranted in practice, however, in that the speaker may take themselves to be the starting point in their pragmatic reasoning chain.

Following the RSA model, the level-1 pragmatic speaker (S_1) used in this chapter maximises the probability that a perspective-taking listener (L_1) will interpret their utterances correctly, using a softmax probabilistic choice rule which describes an approximately optimal decision-maker (see e.g. Franke and Jäger, 2016), as shown in Equation 5.3.

$$P_{S_n}(s|r) \propto e^{\alpha U(s;r)} \tag{5.3}$$

where U stands for the utility of the signal given the intended referent (defined in Equation 5.4 below), and α is a parameter that determines how optimal the speaker is in choosing their signals (α captures the inverse of the error rate in calculating signal utilities; thus higher α leading to more optimal behaviour). In all simulations

reported in this and the following chapter, the optimality parameter α is set to 3.0. This setting was chosen on the basis that it gives pragmatic agents for which the listener has the correct model of the speaker's perspective an advantage for most lexicon types, but only allows them to reach maximum communicative success ($CS = 1.0 - \epsilon$) when using the most informative lexicon type, just like literal agents. Thus, given this parameter setting pragmatic agents have an advantage over literal speakers paired with perspective-taking listeners, but not to the extent that pragmatic agents can reach equal levels of communicative success with different lexicon types. The utility U of a signal is defined as the negative surprisal of the intended referent given the signal for the listener, as shown in Equation 5.4 (following Goodman and Frank (2016)). In order to calculate this utility, a level n pragmatic speaker reasons about a level *n-1* listener, and assumes that this listener shares their lexicon (l_{S_n}) and has the correct model of their perspective ($p'_{correct}$). (Note however that in the iterated learning simulations reported in the next section either or both of these assumptions can be false; the listener may have misinferred the lexicon and/or perspective of their cultural parent.)

$$U_{S_n}(s; r) = ln(P_{L_{n-1}}(r \mid s,\, p'_{correct},\, \ell_{S_n})) \tag{5.4}$$

A level n pragmatic listener L_n then inverts the model of a speaker that is one level below themselves in terms of pragmatic reasoning, S_{n-1}, in order to determine how likely a given signal is to be used by this speaker to refer to the different possible referents. The pragmatic listener L_n does this by simply normalising the speaker's production probabilities ($P_{S_{n-1}}(s \mid r,\ p',\ \ell_{L_n})$; see Equation 5.3) over referents, as shown in Equation 5.2. L_n then interprets utterances according to the resulting probabilities. Note that just like the pragmatic speaker, the pragmatic listener bases their communication behaviour on the assumptions that the speaker shares their lexicon (ℓ_{L_n}) and that their model of the speaker's perspective (p') is correct. As mentioned above, either or both of these assumptions may be unwarranted.

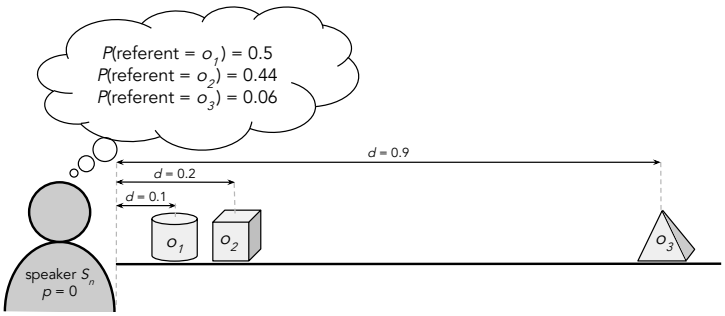
If a level-1 pragmatic speaker (S_1) and level-2 pragmatic listener (L_2) would communicate using the production probabilities as defined in Equation 5.3 and the reception probabilities as defined in Equation 5.2, and in addition both agents would share the same lexicon and the listener would have the correct model of the speaker's perspective, the listener would have perfect knowledge of the speaker's production probabilities. This would give the pragmatic agents an unfair advantage over literal agents, because

the production error parameter ϵ, although having an effect at the bottom of the pragmatic reasoning chain (at the level of the literal speaker), would no longer cause noise in the speaker's utterances that the pragmatic listener cannot predict. Noise in utterance production is therefore reintroduced in all simulations reported below by in each interaction first having the speaker choose a signal with probability equal to $P_{S_n}(s|r)$ as shown in Equation 5.3, and subsequently changing this signal to any of the other signals with a probability equal to ϵ. Although pragmatic learners and listeners have an accurate model of pragmatic speaker's production behaviour, including the probability of added noise, a pragmatic listener cannot predict when and in what direction production errors will happen, and pragmatic speaker-listener pairs will therefore be unable to reach 100% communicative accuracy, even if they use a lexicon of the most informative type.

Figure 5.1 shows an example of a pragmatic speaker and listener (5.1f and 5.1h) compared to a literal speaker and perspective-taking listener (5.1c and 5.1e) as were used in Chapters 3 and 4. In the example shown in this figure, all speakers and listeners share the same lexicon of type 10 (with an informativeness level of 0.53 *ca*), and the listeners have the correct model of the speaker's perspective. Figure 5.1e shows firstly that perspective-taking listeners as used in Chapter 4 already perform one kind of pragmatic inference by taking into account the saliency of the different objects for the speaker (given the context and the listener's model of the speaker's perspective). The lexicon shown in Figure 5.1b is ambiguous: signal s_1 can refer to both referents r_1 and r_2, and s_2 can refer to both r_2 and r_3. A literal listener L_0 would interpret such an ambiguous signal simply by choosing randomly between the referents it is associated with. A literal speaker-listener pair who use the lexicon in Figure 5.1b would thus have a relatively low level of communicative success (0.53 *ca* to be exact; recall that informativeness was defined in Chapter 3 as the communicative accuracy between a literal speaker and a literal listener who both use that lexicon (see Equation 3.7). A perspective-taking listener L_1 as used in Chapter 4 however (see Figure 5.1e) reasons about the literal speaker S_0 by inverting S_0's production probabilities (shown in Figure 5.1d), using the Bayesian inference method of the RSA listener (see Equation 5.2). Thereby the perspective-taking listener L_1 takes into account how likely the speaker is to talk about each of the referents (given the context and the speaker's perspective according to the listener). The perspective-taking listener L_1 in Figure 5.1e has the correct model (p') of the speaker's perspective, and therefore takes into account that

when speaker S_0 uses signal s_1, this signal is more likely to refer to referent r_1 than to r_2, simply because the speaker is more likely to choose r_1 as their intended referent (see the speaker's saliency distribution over referents in Figure 5.1a). Similarly, L_1 knows that when S_0 uses s_2, this is more likely to refer to r_2 than r_3. As shown in Figure 5.1e, the latter inference is stronger than the former because the saliency difference between o_2 and o_3 is bigger than the saliency difference between o_1 and o_2.

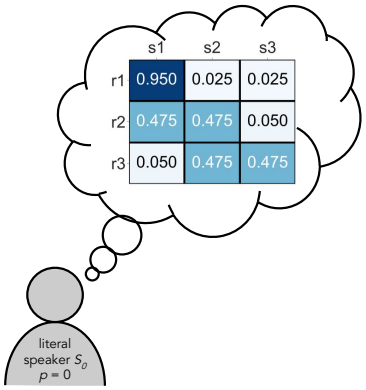
The pragmatic speaker S_1 shown in Figure 5.1f reasons about the perspective-taking listener L_1 and optimises their signal probabilities accordingly. Given that the highest interpretation probabilities for L_1 are along the diagonal of the lexicon matrix (r_1 for s_1, r_2 for s_2, etc.), the pragmatic speaker capitalises on these probabilities by reproducing this pattern in an exaggerated way (see equations 5.3 and 5.4). The level-2 pragmatic listener L_2 (Figure 5.1h) in turn reasons about this pragmatic speaker, and adjusts their interpretation probabilities accordingly; simply an inverted version of the speaker's production probabilities as shown in Figure 5.1g, using the same Bayesian inference procedure as L_1.



(a) Context and corresponding probability distribution over potential referents for speaker with perspective $p = 0$

	s1	s2	s3
r1	1	0	0
r2	1	1	0
r3	0	1	1

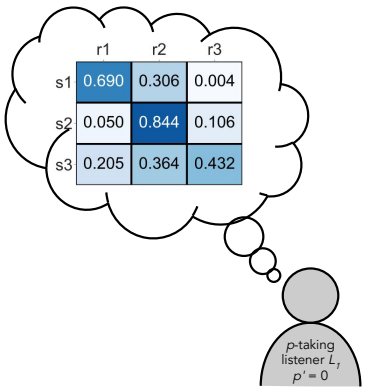
(b) Lexicon used by all agents below (informativeness = 0.53 *ca*)



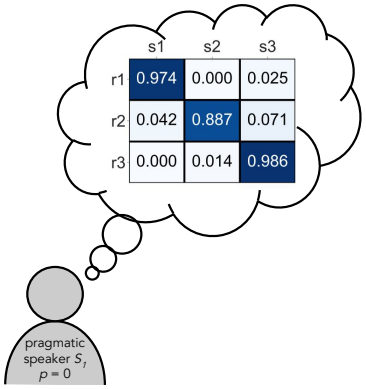
(c) Literal speaker (S_0) with $p = 0$, normalised over referents

	s1	s2	s3
r1	0.475	0.012	0.012
r2	0.211	0.211	0.022
r3	0.003	0.026	0.026

(d) S_0 multiplied with referent saliency



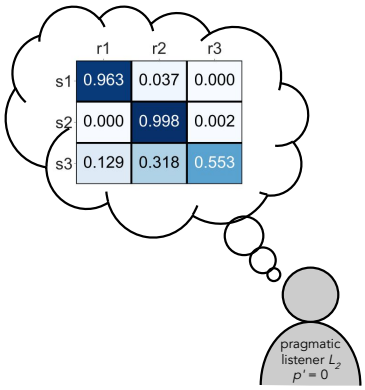
(e) Perspective-taking listener L_1 (reasons about S_0 with $p' = 0$)



(f) Pragmatic speaker S_1 (reasons about L_1, assuming correct lexicon and perspective)

	s1	s2	s3
r1	0.440	0.000	0.060
r2	0.000	0.392	0.053
r3	0.000	0.049	0.007

(g) S_1 multiplied with referent saliency



(h) Pragmatic listener L_2 (reasons about S_1 with $p' = 0$)

Figure 5.1: Diagram illustrating how pragmatic agents compensate for ambiguity in the lexicon. (c) shows the probabilities with which a literal speaker in context (a) with lexicon (b) will produce the different signals for the different referents. (d) shows how a perspective-taking listener who shares the speaker's lexicon and has the correct model of the speaker's perspective would interpret the speaker's utterances (normalised over rows). (e) shows a pragmatic speaker reasoning about the listener in (d), and (f) shows a pragmatic listener reasoning about the speaker in (e). Note that the production probabilities for the speakers in (c) and (e) are normalised over each referent row, and therefore do not reflect the different probabilities with which the speakers will pick each of the objects as their intended referent; this is shown in (a).

From the example in Figure 5.1 we can predict that a pair of a level-1 pragmatic speaker and a level-2 pragmatic listener should have a higher degree of communicative success than the other possible speaker-listener combinations. However, this example only considers one particular lexicon of a relatively ambiguous type, and assumes that the listeners have the correct model of the speakers' perspectives. Figure 5.2 shows the extent to which a pair of a level-1 pragmatic speaker and a level-2 pragmatic listener have an advantage over other possible speaker-listener pairs for the full range of different possible lexicons (grouped by informativeness level). Furthermore, Figure 5.2 also shows the difference in communicative success between listeners who have the correct model of the speaker's perspective and listeners who don't.

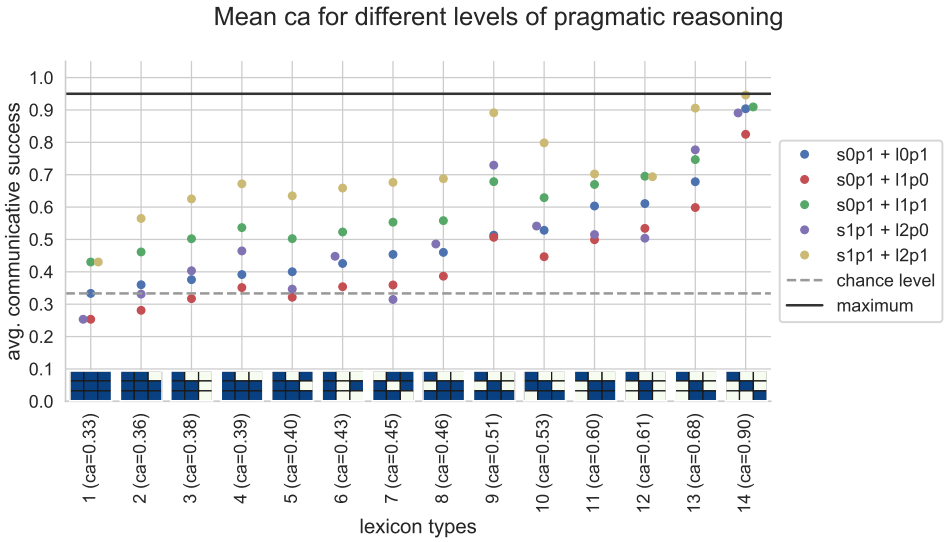


Figure 5.2: Average communicative success for speaker-listener pairs with different pragmatic reasoning levels (ranging from literal-literal to pragmatic-pragmatic), and for listeners having either correct or incorrect model of speaker's perspective (p'). Points show mean over 10,000 randomly generated contexts, and over each possible lexicon within each lexicon type (ranging between 6 and 63 lexicons per type). Communicative accuracy between speaker and listener within each context is calculated exactly (equivalent to the mean of an infinite amount of interactions per context). Pragmatic speaker's optimality parameter α is set to 3.0. Dashed grey line indicates chance level for a world with three possible referents. Solid black line indicates maximum communicative accuracy that can be reached $(1 - \epsilon)$. As in Chapter 4, matrices at bottom of figure represent a single example lexicon for each corresponding informativeness category (i.e. lexicon type). In these matrices, referents are represented by rows and signals by columns (as in Figure 5.1b), and blue squares indicate the presence of an association while white squares represent the absence of one.

Figure 5.2 shows firstly that for all lexicon types, a perspective-taking listener (L_1) who has the correct model of the speaker's perspective performs better at interpreting

the speaker's utterances than a literal listener does. A perspective-taking listener with the incorrect model of the speaker's perspective however, performs systematically worse than a literal listener (in some cases even below chance level, when the lexicon type has low informativeness). This makes sense given that L_1 with incorrect p' adjusts their interpretation probabilities based on an incorrect assumption about the speaker. Furthermore, Figure 5.2 shows that pairs of a level-1 pragmatic speaker (S_1) and a level-2 pragmatic listener (L_2) who holds correct p' outperform the pairs of a literal speaker and perspective-taking listener with correct p' for almost all lexicon types. The only exceptions to this rule are types 1 and 12. In the case of type 1, this is a result of the fact that lexicons of this type are entirely uninformative, which means that there is no difference in expected utility between the signals, no matter what the speaker's intended referent is. Recall that following the RSA model, a signal's utility is defined here as the negative surprisal of the intended referent given the signal for the listener (see Equation 5.4), which is a measure of the information that the listener gains about the intended referent when observing the speaker's utterance (see e.g. Goodman and Stuhlmüller, 2013). In the case of lexicon type 12, the problem is not a lack of difference in expected utility between the signals for each referent per se, but rather a lack of difference in the *ratios* of signal utilities between each referent. Because these ratios are the same across referents (albeit ordered differently), the speaker's maximisation of these expected utilities does not cause any ambiguity resolution. The pragmatic speaker's production probabilities do get shifted relative to those of the literal speaker in such a way that the signal which isn't associated to any referent in the underlying lexicon is made use of. However, because the ratios between the utilities of the other two signals in the lexicon (which *are* associated to either one or two referents) do not differ between the different referents, the pragmatic speaker ends up producing the 'extra' signal with exactly equal proportions for the different referents, rendering this signal as uninformative as it is when being produced by a literal speaker.

Finally, the communicative success of the S_1-L_2 pair where L_2 holds incorrect p' depends on the lexicon type. For most of the lexicon types this pair's communicative success is either just above or just below that of a literal-literal pair. For a few lexicon types however (types 6, 9 and 13), the pair performs slightly better than the S_0-L_1 pair for which L_1 holds the correct p'. In other words, if the pair uses one of these lexicon types, being pragmatic gives agents an advantage over other agent types regardless of whether the listener uses the correct model of the speaker's perspective or

not. On the other extreme, there are also a few lexicon types (7 and 12) for which the pragmatic-pragmatic pair with incorrect p' has even lower communicative success than a literal speaker with a perspective-taking listener who holds incorrect p'. On average, it is more important for level-2 pragmatic listeners to have the correct model of the speaker's perspective than it is for perspective-taking listeners. The difference in communicative success between listeners who hold correct p' and listeners who don't ranges between 0.055 and 0.361 *ca* for pragmatic listeners, with a mean of 0.206 and a standard deviation of 0.071. For perspective-taking listeners in contrast, this difference ranges between 0.085 and 0.192 *ca*, with $\mu = 0.169$ and $\sigma = 0.169$. This makes sense given that for a level-2 pragmatic listener, their model of the speaker's perspective weighs in twice: first in calculating the base rate probabilities for each of the referents being the intended referent given the context ($P(r \mid p')$), and subsequently in calculating the speaker's utterance probabilities (because this pragmatic speaker optimises their utterances with a perspective-taking listener in mind). In other words, when speakers are pragmatic, their perspective has a bigger influence on their communication behaviour than when speakers are literal, because for pragmatic speakers it affects not just how likely they are to choose each referent as their intended referent, but also how likely they are to use the different signals to communicate that intended referent.

5.2.2 Learning from pragmatic agents

Pragmatic learners are implemented in exactly the same way as perspective-taking learners as described in Chapter 3 (Section 3.4.3). They have an accurate model of how pragmatic speakers produce utterances (used to calculate the likelihood of the data given a particular hypothesis about the speaker), but have to infer the speaker's perspective and lexicon (through Bayesian inference). Below I only describe the development and cultural evolution of (populations of) egocentric agents, who start out with a strong bias in favour of the hypothesis that the speaker they're receiving input from shares their perspective on the world ($P(p_{same}) = 0.9$). This is an unhelpful assumption, because in all simulations reported below, the speaker or cultural parent of the learner in fact has the opposite perspective. To correctly infer the perspective of the speaker, such egocentric learners have to overcome their perspective bias using the knowledge they acquire of the speaker's lexicon. The motivation for positing this egocentric bias comes from empirical evidence showing that young children start out

reasoning about other minds from an egocentric perspective, and that this bias diminishes over time (see Birch and Bloom, 2004, for a review); I am therefore particularly interested in the results from this condition.

5.2.3 Iterated learning with pragmatic agents

Iterated learning with pragmatic agents follows exactly the same procedure as was used for literal agents in Chapter 4 (Section 4.2.1). After learning, each individual agent samples a composite hypothesis from their posterior probability distribution, adopts the lexicon part of that hypothesis as their own, and assigns the perspective part of that hypothesis to their cultural parent. Note that thus only lexicons are transmitted culturally over generations; perspectives are determined at birth in such a way that learners always have the opposite perspective to that of their cultural parent. The width of the transmission bottleneck is kept the same as in Chapter 4: 120 observations per learner. Populations start out with the first generation of agents all sharing the same entirely uninformative lexicon, which associates each signal with each of the referents. As in Chapter 4, the question of interest here is under what circumstances these populations evolve a more informative lexicon type over generations, and how this in turn affects agents' success at communicating and inferring each others' perspectives. To answer this question, the current chapter compares three different selection conditions: *No selection*, *Selection for communication* and *Selection on perspective-inference* (each defined in Chapter 4 (Section 4.2.3).

The *Selection on lexicon-learning* condition is not included in the current chapter, because as described in Chapter 4 its effect is solely dependent on how many observations the learner requires to learn each of the different lexicon types, which is highly sensitive to different parameter settings and their interaction. This selection condition was included in Chapter 4 to mirror the *Selection on perspective-inference* condition. However, exploring how robust the results of the *Selection on lexicon-learning* condition are is not within the scope of this thesis. The current chapter focuses only on those selection pressures which have an effect that is expected to be robust against changes in parameter settings that aren't explored in this thesis.

5.3 Learning and evolution of lexicons in pragmatic agents

This section summarises simulation results comparing the development and cultural evolution of lexicons in (populations of) pragmatic agents and (populations of) literal agents (the latter of which are described in more detail in Chapters 3 and 4). The current chapter will focus only on agents with an egocentric perspective bias, which empirical evidence suggests captures something real about children's mindreading development (see e.g. Birch and Bloom, 2004).

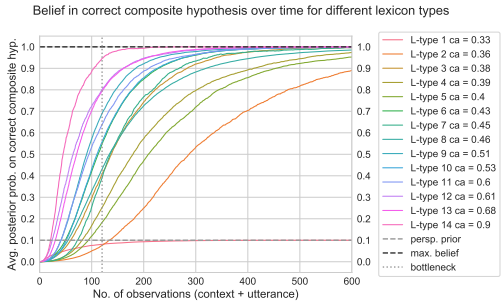
5.3.1 Co-development of lexicon-learning and perspective-taking

In Chapter 3, we saw that lexicon-learning and perspective-learning go hand in hand. Given enough observations made in different contexts, learners who receive data from a literal speaker can correctly infer the speaker's lexicon and perspective. Learning is slowed down when the learner has an egocentric bias and/or when the speaker uses a less informative lexicon, but learners eventually always reach maximum posterior belief in the correct composite hypothesis about the speaker. There are two exceptions to this rule however: (i) if the learner is not able to consider the correct hypothesis about the speaker's perspective (i.e. if the learner's prior assigns zero probability to the correct perspective hypothesis), and (ii) if the speaker uses a completely uninformative lexicon. In the first case the learner will not accumulate any posterior belief in the correct composite hypothesis. This is in fact unsurprising, because the perspective part of this composite hypothesis is simply not considered a possibility by the learner. In the second case the learner is able to acquire the correct lexicon hypothesis (and quite quickly so) but their posterior belief in the correct perspective hypothesis (and therefore in the correct composite hypothesis) will never exceed their prior belief in this hypothesis. This is because a speaker with a completely uninformative (i.e. maximally ambiguous) lexicon produces signals with exactly the same frequency no matter what the context is, and the learner therefore does not have any way into acquiring information about the speaker's perspective.

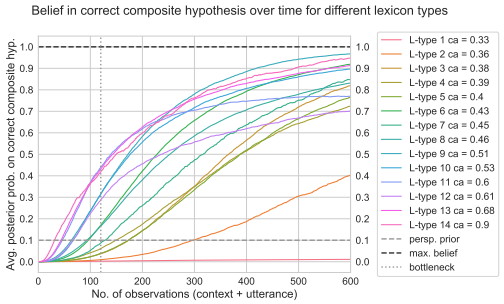
Figure 5.3 compares the learning results of pragmatic learners receiving input from pragmatic speakers to those of perspective-taking learners receiving input from literal speakers (as discussed in more detail in Chapter 3). This figure shows firstly that learning of the composite hypothesis in pragmatic agents is much slower than in literal agents (compare figures 5.3a and 5.3b). When we further compare figures 5.3c and 5.3d

on the one hand and figures 5.3e and 5.3f on the other hand, we see that this is mainly a result of lexicon-learning progressing more slowly in pragmatic agents than it does in literal agents. For perspective-learning in contrast, it depends on the informativeness of the input lexicon whether pragmatic agents are slower or in fact faster at acquiring the correct perspective hypothesis than literal agents are. Pragmatic agents are faster at learning about perspectives when given input from one of the more informative lexicon types, but slower when given input from a less informative lexicon type.

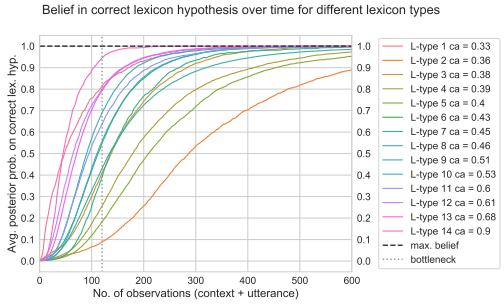
The finding that lexicon-learning is slower in pragmatic agents is a result of the fact that data produced by a pragmatic speaker is more confusable between lexicon types than data produced by a literal speaker. This is discussed in more detail in appendices D and E, but in short, we can define the confusability of two lexicon types as the inverse of the difference between the data produced given one lexicon type and the data produced given another. (That is, the smaller the difference in predicted datasets, the more confusable the two lexicon types are.) The confusability of lexicon types for data produced by a literal speaker ranges from 0.080 to 0.264 (with $\mu = 0.119$ and $\sigma = 0.033$), while for a pragmatic speaker it ranges from 0.108 to 0.622 (with $\mu = 0.188$ and $\sigma = 0.080$). These values are not very interpretable by themselves, given that they are the result of taking the inverse of the sum of the absolute difference between probabilities (see Appendix D for more details). However, the difference in the means and ranges of confusability of lexicon types between the two speaker types tells us that the data produced by pragmatic speakers with different lexicon types is more similar than the data produced by literal speakers with different lexicon types. This explains why pragmatic learners need more observations to infer exactly what lexicon their speaker who provides input is using.



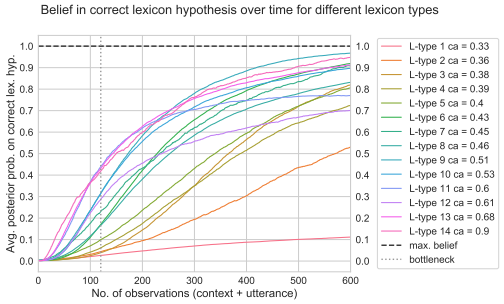
(a) Literal agents; Posterior probability on correct composite hypothesis (i.e. lexicon + perspective)



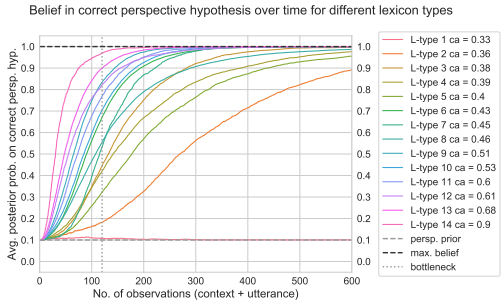
(b) Pragmatic agents; Posterior probability on correct composite hypothesis (i.e. lexicon + perspective)



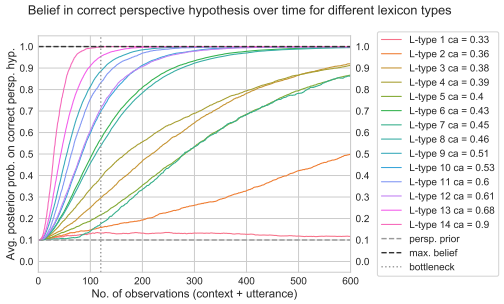
(c) Literal agents; Posterior probability on correct lexicon hypothesis



(d) Pragmatic agents; Posterior probability on correct lexicon hypothesis



(e) Literal agents; Posterior probability on correct perspective hypothesis



(f) Pragmatic agents; Posterior probability on correct perspective hypothesis

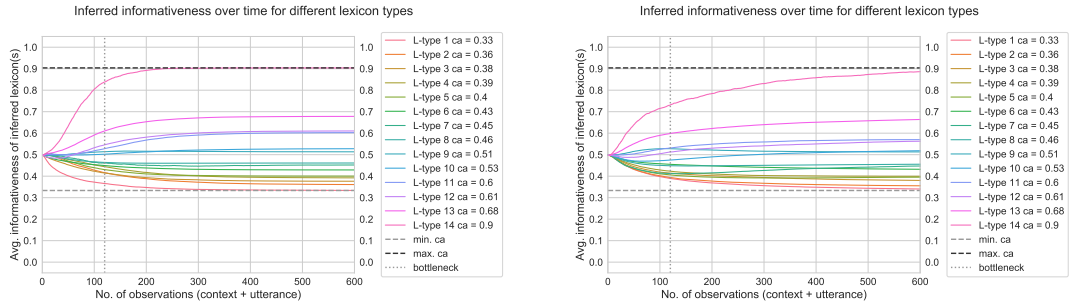
Figure 5.3: Learning curves for pragmatic learners with egocentric bias receiving data from a pragmatic speaker that has the opposite perspective from the learner. Learning curves are separated by input lexicon type (i.e. informativeness class). *ca* levels of different lexicon types range from lowest possible (0.33...) to highest possible (0.90) for lexicon size = 3x3 and error rate $\epsilon = 0.05$. Learning curves show amount of posterior probability assigned to correct hypothesis over time (i.e. number of observed contexts), averaged over 100 independent simulation runs per input language, and subsequently averaged over all languages per informativeness class. Figure *a* shows results for correct composite hypothesis (i.e. lexicon + perspective); figure *b* for correct lexicon hypothesis; and figure *c* for correct perspective hypothesis. Grey dashed line indicates the prior probability assigned to the correct perspective hypothesis. Black dashed line indicates maximum posterior probability that can be reached.

We can also measure how the learner's 'inferred informativeness' develops over time (see also Appendix C). This is obtained by at each time step multiplying the posterior

probability that the learner assigns to each lexicon hypothesis with the informativeness of that lexicon, and summing the resulting values over all lexicon hypotheses, as shown in Equation 5.5.

$$ca' = \sum_{\ell \in \mathcal{L}} P(\ell|\mathcal{D}) \cdot ca(\ell) \tag{5.5}$$

where ca' stands for the informativeness level that the learner has inferred, $\mathcal{L}$ stands for the total space of lexicon hypotheses, $P(\ell|\mathcal{D})$ for the posterior probability of lexicon hypothesis ℓ given data $\mathcal{D}$, and *ca(ℓ)* for the informativeness of lexicon ℓ, measured as the communicative accuracy (*ca*) of the lexicon with itself (see Chapter 3, Section 3.5). This measure of inferred informativeness allows us to see how quickly the learner's belief about the informativeness of the input lexicon reflects its actual informativeness, as shown in Figure 5.4.



(a) Literal agents

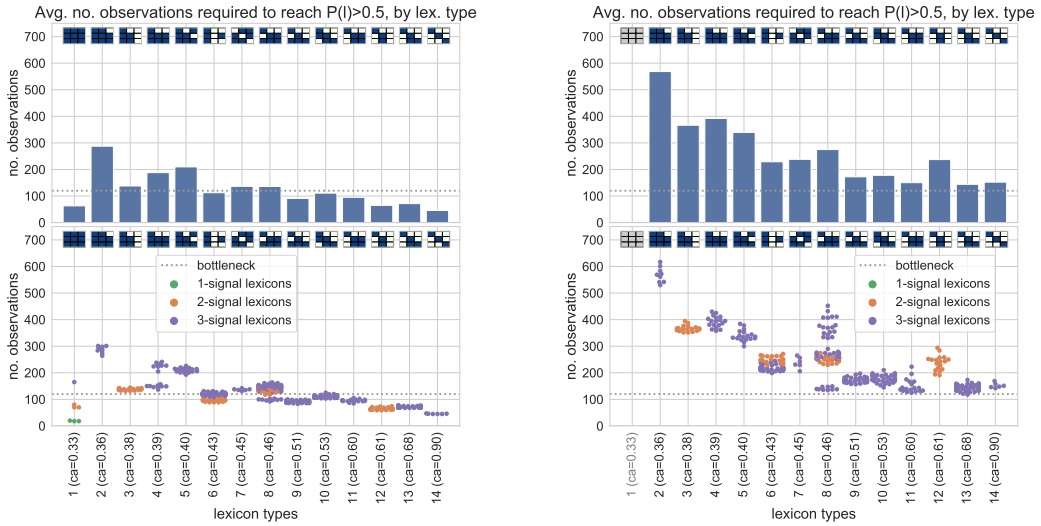
(b) Pragmatic agents

Figure 5.4: Average informativeness of 'inferred' lexicon over time, for all different possible input lexicons, categorised by informativeness. *ca* levels of different lexicon types range from lowest possible (0.33...) to highest possible (0.90) for lexicon size = 3x3 and error rate $\epsilon = 0.05$. Grey dashed line indicates the minimum informativeness that a lexicon can have (equal to chance level for three referents); black dashed line indicates the maximum informativeness a lexicon can have.

Figure 5.4 shows that pragmatic learners take longer to correctly infer the informativeness of the lexicon they're receiving input from than literal learners. This is in line with the finding that pragmatic learners take longer to correctly infer their input lexicon in general, as shown in figures 5.3c and 5.3d. Figure 5.4 furthermore reveals that especially when receiving input from the most informative lexicon type, pragmatic learners have a tendency to underestimate the informativeness of the lexicon they're receiving input from, more so than literal learners. And the other way around, when receiving input from the least informative lexicon type, pragmatic learners have a tendency to *over*estimate its informativeness.

The cause of these findings is well-illustrated by the example in Figure 5.1: data that in the case of literal agents could be produced only by a lexicon with exclusively one-to-one mappings (i.e. a maximally informative lexicon), can in the case of pragmatic agents also be produced by a speaker who uses a less informative lexicon. In addition to this, pragmatic speakers are less tied to the truth-conditional meaning of signals. A pragmatic speaker will occasionally use a signal that is not associated with their intended referent in order to avoid ambiguity elsewhere in the lexicon. A literal speaker in contrast will only ever do so by mistake, which has a relatively low probability of happening: $\epsilon = 0.05$. Thus, the data produced by a pragmatic speaker is (given a limited number of observations) compatible with more different lexicon hypotheses than the data produced by a literal speaker with the same lexicon. This is quantified in appendices D and E, using a measure of confusability of data between different lexicon hypotheses.

Finally, Figure 5.5 shows how many observations pragmatic agents need to reach the $P(l_{correct}) > 0.5$ threshold compared to literal agents. As discussed in Chapter 4 (Section 4.2.2), this threshold gives an indication of how many observations are required for the learner's belief in the correct lexicon hypothesis to exceed their belief in all other possible lexicon hypotheses. It therefore tells us something about which lexicon types are most likely to pass through the transmission bottleneck without being transformed. In line with figures 5.3c and 5.3d, this figure shows that pragmatic learners require about twice as many observations to reach the same level of belief in the correct lexicon hypothesis as literal learners do. There are only a few input lexicons for which pragmatic learners exceed the $P(l_{correct}) > 0.5$ threshold within the bottleneck width of 120 observations (see bottom panel of Figure 5.5b).



(a) Literal agents

(b) Pragmatic agents

Figure 5.5: Average number of observations required for an egocentric learner to reach a posterior belief of $P(\ell) > 0.5$ in the correct lexicon hypothesis when observing only optimal contexts, categorised by informativeness of input lexicon. Top panel shows means over all lexicons within a given lexicon type, where each individual lexicon's value is in turn a mean over 100 independent simulation runs (the same runs of which the results are shown in figures 5.3 and 5.4 above). Bottom panel shows individual lexicons (same means over 100 simulation runs), colour-coded by how many signals they make use of. Grey dashed line shows number of observations used for all simulations reported below (i.e. the transmission bottleneck; 120 observations). The number of observations required to reach the $P(\ell) > 0.5$ threshold is not depicted for lexicon type 1 for pragmatic agents, because for pragmatic learners it is impossible to distinguish between the different lexicons comprised in this lexicon type. (I.e. the production probabilities for pragmatic speakers look exactly the same for each of the different lexicons of type 1.)

There is also a difference between pragmatic and literal agents in terms of how the number of signals that is made use of by the lexicon affects the number of observations the learner needs to acquire it. While for literal agents lexicons that make use of fewer signals are learned faster (because there are fewer mappings to learn), this is not the case for pragmatic agents. This is a result of the fact that if a lexicon in its basic form with binary referent-signal associations makes use of less than all three available signals, a pragmatic speaker will nevertheless put the remaining signals to use in order to resolve ambiguity. For most of these lexicon types this causes pragmatic agents to have higher communicative success than literal agents (see lexicon types 3, 6 and 8 in Figure 5.2). However, it makes it harder for learners to acquire the lexicon, because it is not possible for them to quickly disregard all lexicon hypotheses that use all three

signals (of which there are far more than there are of those that use only one or two of the signals).

The number of observations required to reach the $P(\ell) > 0.5$ threshold is not shown for lexicon type 1 for pragmatic agents, because it is impossible for pragmatic learners to distinguish the different lexicons that are comprised in this type. This is a result of how pragmatic speakers determine their signal probabilities given referents. As described in Section 5.2.1, pragmatic speakers reason about a perspective-taking listener, and this perspective-taking listener in turn reasons about a literal speaker and inverts the production probabilities of this literal speaker to yield a set of interpretation probabilities. Because each of the signals in each of the lexicons of type 1 are uninformative, the perspective-taking listener simply assigns the speaker's referent choice probabilities given the context to each of the referents, regardless of which signal is used. Therefore, as far as the perspective-taking listener is concerned, each of the signals is equally (un)useful for lexicons of this type, which means that the pragmatic speaker has nothing to optimise over. Thus, a pragmatic speaker who uses any of the lexicons of type 1 will always end up assigning exactly equal production probabilities to each of the signals for each of the referents. This means that the pragmatic speaker's production behaviour will look like that of a literal speaker who uses the lexicon which associates each of the signals with each of the referents, regardless of which signal(s) the speaker's underlying lexicon actually makes use of. Thus, the data produced by a pragmatic speaker will be identical for each of the lexicons of type 1, leading to these lexicons being unlearnable for a pragmatic learner.

From the results shown in Figure 5.5, we can conclude that, keeping the bottleneck width constant, transformations will be more likely in iterated learning chains with pragmatic agents than they are with literal agents. How this, in combination with pragmatic agents' different development and communicative success, affects the cultural evolution of lexicons in populations of pragmatic agents is explored in the next section.

5.3.2 Pragmatic agents can be successful communicators and perspective-takers despite ambiguous lexicons

This section presents simulation results that show how the different development and communicative success of pragmatic agents affects the cultural evolution of their lexicons. As in Chapter 4, different selection conditions are compared: *No selection*,

Selection for communication and *Selection on perspective-inference*. As described in more detail in Chapter 4 (Section 4.2.3), agents in the *Selection for communication* condition are selected to become cultural parents with probability proportional to how successful they are at interpreting the utterances of their cultural parent. In the *Selection on perspective-inference* condition in contrast, agents are selected proportional to how much posterior probability they assign to the correct hypothesis about their cultural parents' perspective.

Figure 5.6 shows how the average informativeness of selected lexicons changes over generations in populations of pragmatic agents compared to populations of literal agents. This figure shows firstly that the maximum convergence point across selection conditions differs between literal and pragmatic populations. The convergence point was defined in Chapter 4 as the generation from which onwards the variation in average informativeness of the population remains within a range of 0.1 *ca* for a minimum of 50 consecutive generations, for each individual simulation run. Because the convergence point differs between selection conditions, the highest one is applied uniformly across selection conditions within a given population type (i.e. literal or pragmatic). This convergence point was then used to determine the burn-in period that is discarded in measures of populations' success and the equilibrium distribution over lexicon types (reported below).

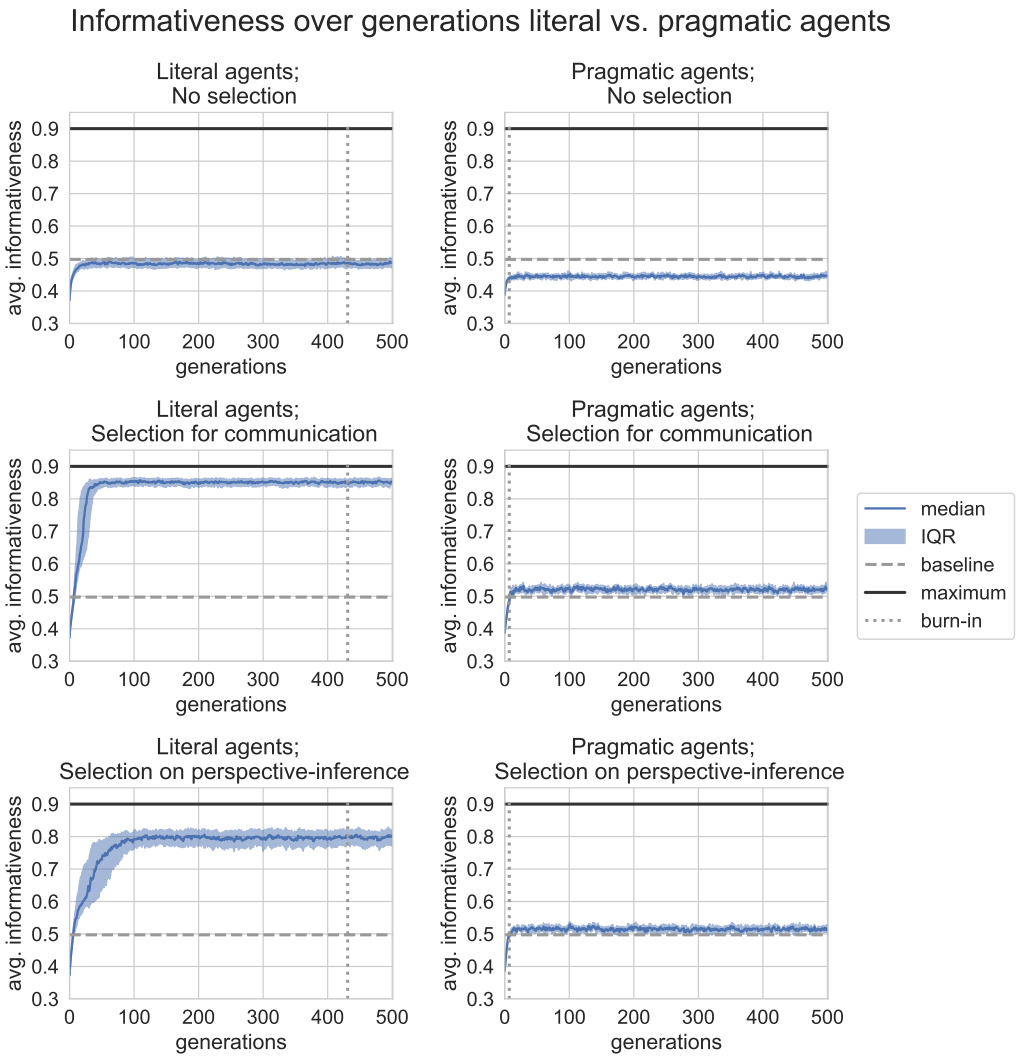


Figure 5.6: Average informativeness of lexicons in literal and pragmatic populations under different selection pressures. Solid blue line shows median and shaded area shows interquartile range over independent simulation runs (100 runs for literal populations and 25 runs for pragmatic populations). Dashed grey line shows the baseline informativeness one would expect if agents are picking lexicons at random. Minimum informativeness is 0.33... *ca* (given lexicon size 3x3), which is the starting point for all populations. Maximum informativeness is 0.90 *ca* (given production error $\epsilon = 0.05$), which is indicated by the solid black line. Dotted grey line indicates the final generation of the burn-in period which is discarded for calculating the populations' success and the equilibrium distribution over lexicon types (both reported below).

Figure 5.6 further shows that in populations of pragmatic agents, the average informativeness of the population's lexicons in the *No selection* condition does not converge to the level that would be expected if agents are picking lexicons at random (as is the case in literal populations). Instead, the average informativeness in pragmatic popu-

lations remains lower. As will be shown in Figure 5.8 below and discussed in more detail there, this is a result of pragmatic agents in this condition selecting the less informative lexicon types more often than the more informative ones. In the *Selection for communication* and *Selection on perspective-inference* conditions, average informativeness increases relative to the *No selection* condition, but much less so in pragmatic populations than it does in literal populations. The average informativeness in pragmatic populations only just exceeds the level that would be expected if agents pick their lexicons at random. However, Figure 5.8 below shows that random selection is not what is happening in this condition.

Figure 5.7 shows the average success that these populations reach in terms of communicating with and inferring the perspectives of their cultural parents after convergence. This figure reveals that despite the relatively low levels of informativeness of the lexicons that the pragmatic populations converge on, they nevertheless reach substantially higher levels of communicative and perspective-inference success under the two selection conditions than they do in the *No selection* condition. The increase in success that is reached under selection is not quite as big in the pragmatic populations as it is in the literal populations, but the pattern of results looks the same. Despite a relatively small gain in average informativeness, the pragmatic populations reach higher levels of success at both communication and perspective-inference under a selection pressure for communication, and the same holds under

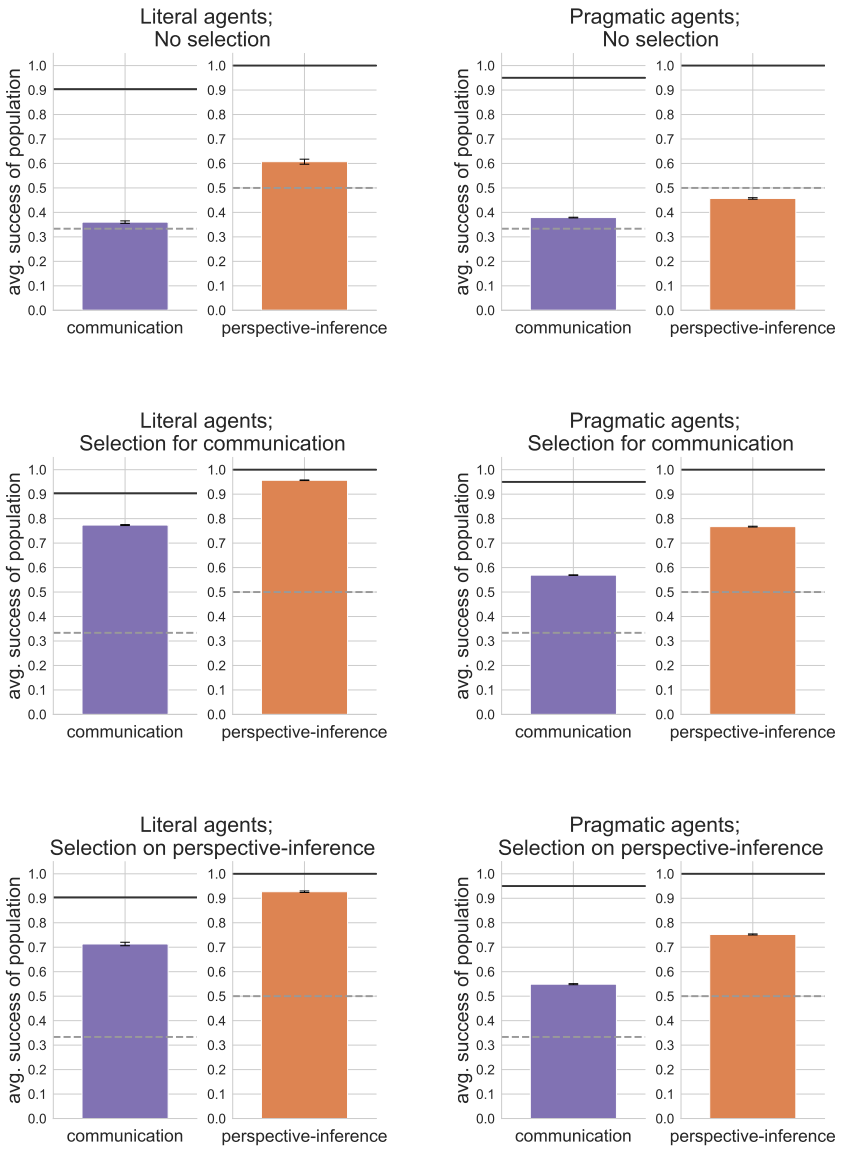


Figure 5.7: Populations' average success at interpreting cultural parents' utterances (communication) and inferring cultural parents' perspective (perspective-inference), after convergence. Each of these measures is calculated independently from agents' fitness as used for selection. Graphs for literal populations show grand means over 100 runs, 69 generations (i.e. a burn-in period of 431 generations is excluded) and 100 agents per generation. Thus each bar shows the grand mean success of 690,000 individual agents. Graphs for pragmatic populations show grand mean success over an equal number of individual agents, but taken over 25 runs and 276 generations (i.e. a shorter burn-in period of 224 generations is excluded). Dashed grey line indicates chance level and solid black line indicates ceiling. Note that the ceiling for communication success differs slightly between literal and pragmatic populations. Given $\epsilon = 0.05$, it lies at 0.90 *ca* for literal agents and at 0.95 *ca* for pragmatic agents.

Finally, Figure 5.8 shows the distribution over lexicon types after convergence. In line with the findings on average informativeness, this figure shows that in the absence

of any selection pressure, pragmatic populations have a stronger tendency to pick less informative lexicon types over more informative ones than literal populations do. (Lexicon types 1, 2 and 7 are selected with proportions more than one standard deviation above the mean, while types 13 and 14 are selected with proportions less than one standard deviation below the mean.) This is in line with the individual learning results described in Section 5.3.1. Specifically, as discussed above, Figure 5.4 shows that especially when given input from the most informative lexicon types, pragmatic learners have a tendency to underestimate the informativeness of their input lexicon, more so than literal learners do. Thus, data produced by a pragmatic speaker with a relatively informative lexicon can, given a limited number of observations, easily be mistaken for data produced by a pragmatic speaker with a less informative lexicon (i.e. one that contains more ambiguity). This is reflected in the measure of *confusability* of lexicon types which is discussed in appendices D and E. Specifically, Appendix E shows that for most lexicon types, the data they produce is most easily confused with (i.e. most similar to) data produced by lexicon types that are at the lower end of the informativeness spectrum. One would therefore expect a population of pragmatic agents to transition into the less informative lexicon types more easily than they would transition out of them.

Also in line with the findings on average informativeness described above, pragmatic populations in the *Selection for communication* and *Selection on perspective-inference* conditions converge on a much more varied set of lexicon types than literal populations do. In both selection conditions, the lexicon types that stand out as particularly well-represented in the pragmatic populations compared to the literal populations are types 9 and 13. (Under both selection pressures, these are the only two lexicon type which are selected with a proportion higher than one standard deviation above the mean[5]). This finding requires a different explanation for the two different selection pressures however, given that agents' probability of becoming a cultural parent depends on different (albeit related) attributes.

[5]For lexicon type 9 this proportion also exceeds two standard deviations above the mean, again in both selection conditions.

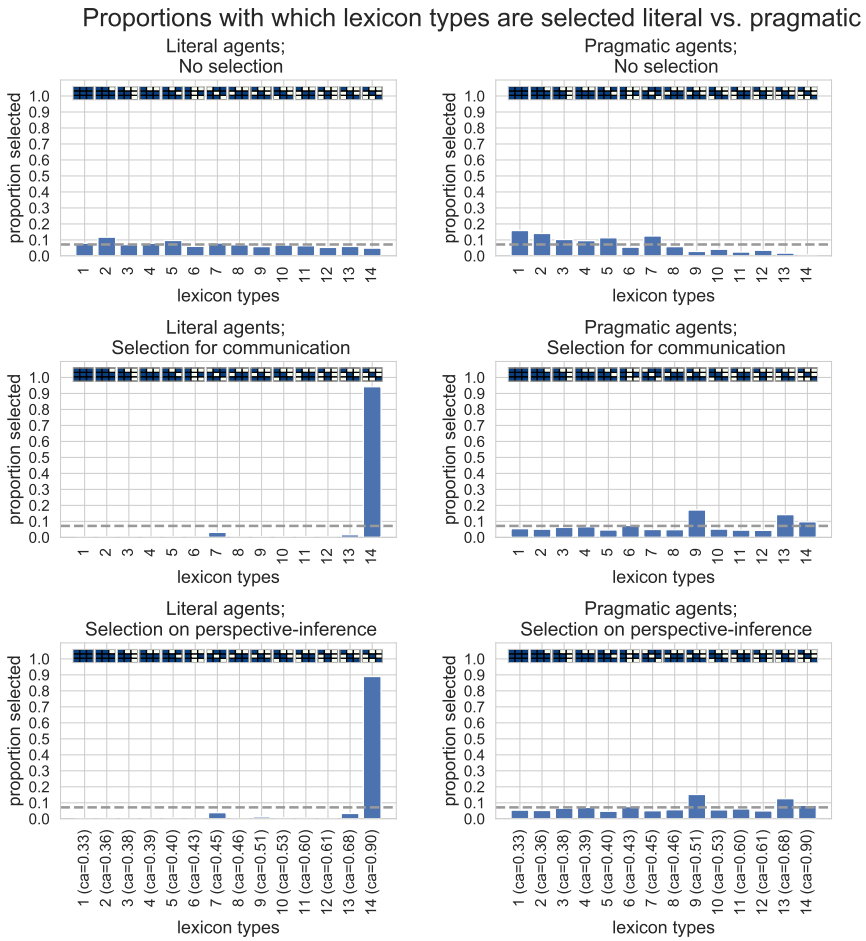


Figure 5.8: Average proportions with which agents select the different lexicon types after convergence. Graphs for literal populations show grand means over 100 runs, 69 generations and 100 agents per generation (690,000 lexicon-selection events in total). Graphs for pragmatic populations show grand means over an equal number of lexicon-selection events, but taken over 25 runs and 276 generations (i.e. a shorter burn-in period of 224 generations is excluded). These grand means are further averaged over the total set of possible lexicons within each lexicon type (ranging between 6 and 63 lexicons per type). The x-axis shows all informativeness categories that exist for 3x3 lexicons, with informativeness levels ranging from lowest possible (0.33... *ca*) to highest possible (0.90 *ca*), given error rate $\epsilon = 0.05$. Dashed grey line shows the baseline distribution over lexicon types that would be expected if agents select lexicons at random.

For the *Selection for communication* condition, the finding that lexicon types 9 and 13 are overrepresented is explained by the fact that together with lexicon type 14, these are the lexicon types that lead to the highest levels of communicative success for pragmatic agents, as shown in Figure 5.2. This is the case both for pairs of pragmatic agents in which the listener has the correct model of the speaker's perspective, and for pairs of pragmatic agents in which the listener has the *in*correct model of the speaker's

perspective. For the *Selection on perspective-inference* condition, the finding that lexicon types 9 and 13 are overrepresented is explained by the fact that these types provide relatively good input for learning about perspectives. As Figure 5.3f shows, the three lexicon types for which learning about the speaker's perspective happens most quickly are types 9, 13 and 14. In addition to these factors, another advantage for lexicon types 9 and 13, which holds across both selection conditions, is that they stand out in terms of requiring (at least on average) fewer observations to exceed the $P(l_{correct}) > 0.5$ threshold than the lexicon types that surround them in the informativeness ranking (in both directions), as shown in Figure 5.5. This indicates that these lexicon types are more likely to pass through the transmission bottleneck unharmed compared to other lexicon types of similar informativeness levels.

5.4 Discussion

In this chapter, the model of speaker and listener behaviour as presented in Chapters 3 and 4 was enriched such that speakers tune into the fact that listeners will take their perspective during communication. Such pragmatic speakers assume that the listener will take their perspective when interpreting utterances, and optimise their utterances to maximise the information that such a perspective-taking listener would gain about their referential intention given the utterance. The developmental results obtained with this pragmatic version of the model showed that overall, receiving input from a pragmatic speaker slows learning down. This is because the data that different pragmatic speakers with different lexicons produce is on average more confusable than the data produced by literal speakers. This in turn is a consequence of the fact that pragmatic speakers are less tied to the truth-conditional meaning of signals when choosing their utterances. They are happy to occasionally use a signal that is not associated with the intended referent according to their lexicon, or, the other way around, to *not* use a signal that *is*, if this helps resolve ambiguity.

When it comes to learning about the speaker's perspective, the picture is a little bit more nuanced. Just as with lexicon-learning, pragmatic learners are at a disadvantage compared to literal learners when receiving input from the less informative lexicon types. However, when receiving input from a restricted set of higher-informative lexicon types, pragmatic learners are instead faster than literal learners at inferring the speaker's perspective. This is presumably a result of the fact that a pragmatic speaker's

communication behaviour is influenced by his perspective twice: both in terms of how likely he is to pick different intended referents and in terms of how likely he is to produce different signals to communicate such an intended referent. The latter is not the case for a literal speaker, and is a consequence of the fact that the pragmatic speaker reasons about a perspective-taking listener.

Aside from needing more observations to achieve the same level of knowledge about the combination of the speaker's lexicon and perspective (i.e. the correct 'composite hypothesis'), pragmatic agents' learning is not really qualitatively different from that of literal agents. The informativeness of a lexicon is still a rough predictor for how quickly the learner will acquire it, and only in the case of the uninformative lexicon type will the learner never be able to accumulate more than their prior belief in the correct hypothesis about the speaker's perspective.

What changes more qualitatively when agents are pragmatic instead of literal communicators, is the result of iterated learning. In the absence of any selection pressure, pragmatic populations converge on sets of lexicons that have a lower average informativeness level than what literal populations converge on. When it comes to communicative success, this lower level of informativeness is compensated for by the agents' pragmatic abilities, but when it comes to perspective-inference, pragmatic populations in this condition are less successful than literal populations.

In the *Selection for communication* and *Selection on perspective-inference* conditions, the gain in informativeness relative to the *No selection* condition is not nearly as big for pragmatic populations as it is for literal populations. However, pragmatic populations under either selection pressure nevertheless improve considerably in their communicative and perspective-inference success, despite only a small gain in informativeness. Under selection for communication, this is a result of pragmatic populations converging on lexicon types that maximise communicative success when communication is pragmatic. Under selection on perspective-inference, it is a result of pragmatic populations converging on lexicon types that maximise the speed at which perspectives are inferred. Because these factors are related (i.e. both are a result of better inference of the speaker's intended referents) these two criteria are fulfilled by the same three lexicon types, and these are thus the most selected lexicon types in both selection conditions. Moreover (and again for the same underlying reasons), these same three lexicon types have a relatively high chance of being transmitted accurately, which adds to their competitiveness in both selection conditions.

What is not qualitatively different between pragmatic and literal populations, is that a selection pressure that selects for either only communication or only perspective-inference, causes an increase in success not just in the skill that is selected for, but also in the other skill. In the case of pragmatic agents this is not a result of populations simply converging on the most informative lexicon type, as is the case in literal populations, but more specifically of the fact that the criteria that increase agents' fitness under the two selection pressures (maximising the success of pragmatic communication on the one hand, and maximising fast and accurate perspective-inference on the other hand) are satisfied by the same three lexicon types.

In sum, if agents don't just have the ability to learn about others' perspectives and use this in utterance comprehension, but on top of this also have the ability to reason about other agents as being such perspective-takers when determining what utterances to produce, this takes pressure off the lexicon in terms of how strictly unambiguous it needs to be. Granted, part of the reason why pragmatic populations who are exposed to a selection pressure end up with a more varied set of lexicons and a lower level of average informativeness, is simply that from data produced by a pragmatic speaker it is generally harder to make out which lexicon he's using than it is from data produced by a literal speaker. More specifically, the more informative lexicons tend to be confused with the less informative ones, leading to transformations towards lower levels of informativeness to be more likely than transformations in the other direction. However, the other part of the reason is that when agents are pragmatic communicators, there are more different lexicon types which support successful communication and perspective-inference compared to when agents are literal communicators, and one of these lexicon types is quite ambiguous.

Under selection for communication, this relatively ambiguous lexicon type can do quite well by virtue of pragmatic communicators' ability to resolve ambiguity by reasoning about each other. Under selection on perspective-inference, this result is a little less straightforward, and has to do with the fact that the formalisation of pragmatic communication presented here leads to there being a restricted set of lexicon types which cause pragmatic speakers' utterances to reveal more about their perspective than the utterances of literal speakers with the same lexicon type. That is, in the model of pragmatic communication used here, pragmatic speakers' perspectives do not only affect how likely they are to talk about different referents, but also how likely they are to choose different utterances to communicate those referential intentions. Thus,

pragmatic speakers' communication behaviour is affected by their perspective twice.

This in turn is a result of the somewhat awkward assumption that a pragmatic speaker adapts his utterances not based on his knowledge about the listener's perspective, but instead on the knowledge that the listener will take his own (i.e. the speaker's) perspective during comprehension. The chain of pragmatic reasoning in this model thus bottoms out at a literal speaker instead of a literal listener. This assumption is a result of the fact that this thesis uses Bayesian inference as a model of learning, which requires the learner to observe a bunch of data and update their beliefs in the different possible generative models accordingly. If we want a Bayesian learner to simultaneously learn a lexicon and learn something about the speaker's perspective on the world, this means both these aspects have to somehow influence what the data looks like, and the learner needs to have an accurate model of how this data is generated. In the case of literal agents therefore, the learner is essentially a perspective-taking listener who reasons about a literal speaker. And to make the implementation of pragmatic agents as similar as possible to that of literal agents, the current model simply builds an extra layer of reasoning for each party (i.e. speaker and listener) on top of this baseline model of literal agents.

It is possible to instead implement pragmatic communication as a speaker who reasons about a literal listener whose interpretation of utterances is affected by her (i.e. the listener's) perspective on the world. (That is, to have the chain of pragmatic reasoning bottom out at a literal listener.) However, in that model the learner's task would change from simultaneously inferring the speaker's lexicon and perspective back to only inferring the speaker's lexicon, or to simultaneously inferring the speaker's lexicon and whether the speaker is a literal or pragmatic communicator. The latter is an interesting model in its own right, but would not allow for a direct comparison of individual learning and iterated learning results between literal and pragmatic agents, as was the aim of this chapter. As Smith et al. (2013) point out however, if a learner was receiving input from a pragmatic speaker, the rational thing to do for this learner would actually not be to try and learn the speaker's underlying lexicon, because (as mentioned above) a pragmatic speaker's production behaviour is not tied to truth-conditional meaning. That is, the pragmatic speaker can use signals that are not actually associated with his intended referent according to his lexicon.

This brings us to a second odd assumption of the model presented in this chapter however: the pragmatic learner *does* in fact attempt to learn a lexicon of binary

mappings from a level-1 pragmatic speaker. Smith et al. (2013) solve this paradox by having learners assume that there is one true lexicon in the population and that all other agents know what it is. In the current model, the paradox is solved by the fact that the pragmatic reasoning chain bottoms out at a literal speaker who has the same lexicon and perspective as the pragmatic speaker. This level-0 literal speaker thus represents the truth-conditional utterance behaviour that the pragmatic speaker ultimately bases their utterances on. The learner in the model presented in this chapter has an accurate model of how this pragmatic speaker produces their utterances, and can therefore infer the speaker's underlying lexicon and perspective in a rational way, by performing Bayesian inference on the utterances of a speaker whose production behaviour is ultimately determined by a lower-level model of themselves.

To conclude, this chapter showed that pragmatic agents under either of the two selection pressures can reach fairly decent levels of success at communicating and inferring others' perspectives without the need for evolving a completely unambiguous lexicon. This is relevant because natural languages today are successful despite containing a substantial amount of ambiguity, and because the humans who use these languages are pragmatic communicators. However, it is as yet unclear under what circumstances such pragmatic communication would evolve. In other words, under what circumstances does it pay off to be a pragmatic communicator rather than a literal one? If a well-functioning system of unambiguous, literal communication already exists, under what circumstances could pragmatic communication nevertheless take over? This question is investigated in the next chapter, in which single pragmatic 'mutants' are introduced into populations of literal agents who have already converged on a lexicon type, in order to explore how likely pragmatic agents are to take over.

Chapter 6

Gene-culture co-evolution of pragmatic ability and lexicons

In the previous chapter we saw that being pragmatic allows populations to be fairly successful at communicating and inferring each others' perspectives even if they don't converge on an entirely unambiguous lexicon type. Natural languages today contain a substantial amount of ambiguity, and Piantadosi et al. (2012) argue that this is an inevitable and in fact desirable feature of languages if their users are pragmatic. Specifically, Piantadosi et al. (2012) show that ambiguity is the outcome of a trade-off between two communicative pressures which they take to be inherent to all communication systems: clarity and ease.

Firstly, using an information-theoretic argument, Piantadosi et al. show that if context is informative about meaning (and listeners have the ability to take that context into account during interpretation), the most efficient communication system will always be an ambiguous one (i.e. one from which the intended meaning could not be recovered with 100% certainty if context was not available). An *un*ambiguous communication system would be redundant in the sense of providing more information than is strictly necessary to recover the communicative intention, and therefore inefficient. Secondly, Piantadosi et al. argue that an ambiguous communication system allows for the reuse of elements (words and sounds) that are easier to produce and understand. This second argument predicts that linguistic units that require less effort (either in production or comprehension) should show more ambiguity. Piantadosi et al. test this prediction empirically using corpus data from three different West Germanic languages. Specifically, they demonstrate that words which are (i) shorter, (ii) higher-frequency,

or (iii) less phonotactically surprising (all features that they take to reflect greater ease of use) have both more meanings (higher rates of homophony) and more senses (higher rates of polysemy), and that units at the syllable-level show a similar pattern. Taken together, Piantadosi et al. argue that ambiguity is a desirable feature of a communication system when context is informative about meaning. In other words, as long as ambiguity in a communication system is not too costly (for instance because it is compensated for by users' pragmatic ability), it is in fact a useful feature, making the system more efficient.

The lexicons that evolve in populations of pragmatic agents as shown in Chapter 5 look more similar to natural languages than the ones that evolve in literal populations as shown in Chapter 4, in the sense that they seem to be as ambiguous as is possible without it being costly to populations' communicative and perspective-inference success. In line with the arguments of Piantadosi et al., this is a result of agents being pragmatic rather than literal, because this pragmatic ability allows learners to accumulate a sufficient amount of information about their cultural parents' intended referents and perspective, while receiving input from a more ambiguous lexicon.

However, the results presented in Chapter 5 do not tell us anything about the 'evolvability' of such pragmatic abilities. That is, assuming that pragmatic reasoning is costly and requires a higher degree of cognitive sophistication (be it as the result of biological or as the result of cultural evolution), under what circumstances would such pragmatic abilities be selected for? This question can be rephrased as a question of evolutionary stability: Under what circumstances is an invading pragmatic 'mutant' likely to take over when it enters a population of literal agents? This is the question that the current chapter is concerned with. In Section 6.1 I will briefly review similar models of gene-culture co-evolution. In Section 6.2 I will describe the simulations with which the question of pragmatic evolvability is tackled here, followed by a description of the resulting findings in Section 6.3. Finally, in Section 6.4 I will discuss these findings in relation to the wider topic of this thesis.

6.1 Review of models of gene-culture co-evolution in language

In Chapter 4 (Section 4.1.1) I reviewed previous iterated learning models with Bayesian agents, which showed that two different methods of hypothesis selection — sampling

and maximum a posteriori (MAP) selection — result in different outcomes of iterated learning. Specifically, Griffiths and Kalish (2007) and Kirby et al. (2007) showed that sampling causes the populations' stationary distribution over languages to converge to the individual learners' prior bias, while MAP selection causes the prior to be amplified, resulting in overrepresentation of the language type that is favoured by the prior.

Smith and Kirby (2008) asked which of these two hypothesis selection methods would be favoured by biological evolution if agents' fitness depends on their ability to communicate with others (where communicative success is defined as the probability that an agent shares their language with a randomly selected peer from the same generation). Smith and Kirby showed firstly that when comparing separate populations of sampling learners and MAP learners, biological selection for communication should favour MAP learners: the latter reach higher levels of within-population communicative success. This is a result of the fact that MAP selection leads to amplification of learners' prior bias, so that given the same prior strength, MAP populations end up with a higher proportion of the language type favoured by the prior than sampling populations do. This in turn leads to MAP populations having a higher probability of two randomly selected agents sharing the same language (after convergence) than sampling populations.

However, as Smith and Kirby (2008) point out, this analysis only shows which hypothesis selection strategy and prior bias are objectively best for populations' communicative success. It does not tell us how likely these features are to actually evolve in a population. The latter question is similar to the question that the current chapter is concerned with. In order to answer this question, Smith and Kirby assume that a population consists of two subpopulations, where a subpopulation consists of agents who all share the same combination of hypothesis selection strategy and prior bias (*learning behaviour* for short). Smith and Kirby further assume (following Griffiths et al. (2008)) that the probability of each of the languages in a subpopulation is equal to its probability in the stationary distribution (given that particular subpopulation's learning behaviour). These assumptions allowed Smith and Kirby to calculate the exact relative fitness of one particular learning behaviour compared to another. Maynard Smith and Price (1973) showed that if the relative fitness (in this case defined as the relative communicative accuracy) of a single learner A compared to a large and homogeneous population of learners of type B exceeds 1.0, learning behaviour A has a reproductive advantage over learning behaviour B. If we further assume (as Smith and Kirby do)

that learning behaviours are genetically inherited, this means that the further the relative fitness of a given learning behaviour exceeds 1.0, the more likely it is to spread through the population and ultimately reach fixation (i.e. a state where this is now the only genotype present in the population). If the relative fitness of A with respect to B is less than 1.0, strategy A will be selected against, and is therefore unlikely to spread through the population. Finally, if the relative fitness of the two learning behaviours is exactly 1.0, the two strategies are selectively neutral, and their frequencies will change according to genetic drift.

Using this analysis, Smith and Kirby (2008) showed that for populations of sampling agents the evolutionarily stable strategy (as defined by Maynard Smith and Price, 1973) is to have the strongest possible prior bias, whereas for populations of maximisers, any bias strength is an evolutionarily stable strategy, as long as there is *some* bias. The latter is a result of the fact that cultural transmission with MAP selection amplifies the effect of individual learners' bias over generations, such that ultimately the exact strength of the bias is in fact 'masked' (i.e. different strength of the bias all result in the same stationary distribution over languages). This masking of the bias strength as a result of MAP selection in turn causes the strength of the bias to be shielded from selection: if a subpopulation of weakly-biased agents ends up with the same selection of languages as a subpopulation of strongly-biased languages, neither of these bias strengths has an evolutionary advantage over the other. Thus, bias strength will not be the subject of selection in populations of maximisers (except that being completely *un*biased will be selected against). This has implications for theories about the co-evolution of language (as a product of cultural evolution) and the language faculty (as a product of biological evolution). If MAP selection captures something real about how learners acquire language, this would lead to an opaque relationship between language as the product of cultural evolution, and the underlying language faculty, thereby ruling out positive selection for a strongly constrained language faculty (equivalent to a strong prior bias in this model).

Furthermore, using the same method of analysis, Smith and Kirby (2008) asked which hypothesis selection strategy is more likely to evolve, sampling or MAP. This analysis showed that MAP selection is an evolutionarily stable strategy (i.e. has relative fitness > 1.0 compared to sampling) no matter how strong the prior bias is, while sampling is not. No matter the bias strength, MAP selection increases the probability that the most likely language will be learned (given the same set of data) compared to

sampling. Therefore, assuming, as Smith and Kirby (2008) did, that agents' success depends on how many of their peers they are able to communicate with, MAP is always the best hypothesis selection strategy. Thus, selection on agents' ability to communicate with their peers leads to MAP selection, and this in turn leads to bias strength to be shielded from selection. Finally, Smith and Kirby show that if we further assume that having a strong bias comes at a certain cost (e.g. because it requires additional, more restrictive cognitive machinery), selection will favour the weakest possible bias.

Similar results were found by Thompson et al. (2016), who additionally showed that MAP selection causes not just masking of the bias strength for selection, but also unmasking of the bias when it first enters the population. That is, when an individual with a very weak bias enters a population of unbiased individuals, MAP selection will amplify the effect of that weak bias on the population level, and thus make it visible to selection. Subsequently however, MAP selection also masks further differences in bias strength from selection, as explained above, making the evolution of a strong innate bias unlikely.

6.2 A model of gene-culture co-evolution of lexicons and pragmatic ability

As mentioned above, the goal of the current chapter is to explore under which circumstances pragmatic agents have an evolutionary advantage over literal agents. The simplest way of addressing this question is to model pragmatic ability as a genetically inherited skill (i.e. one gene with two alleles: 'literal' and 'pragmatic') and to look at under what circumstances the pragmatic allele can invade a population of literal agents. As discussed above, the evolutionary advantage of a particular allele (i.e. communication type in this case) can be captured in its relative fitness compared to another allele. However, in the simulations presented in this chapter, the assumption is that biological and cultural inheritance are decoupled. That is, a new agent that enters the population can receive her communication type gene from one agent, while receiving her linguistic input from another. Thus, each new agent in this model has a biological and a cultural parent, which are selected independently from each other, based on their fitness under the selection condition that they are exposed to.

If biological and cultural inheritance were not decoupled, this would make it easier for the pragmatic mutant allele to invade the population, because each new learner who

inherits this allele would also automatically receive input from a pragmatic speaker. The reason this would make it easier for the pragmatic allele to invade is that the current model further assumes that learners do not infer the communication type of the parent through learning, but instead simply assume that whatever their own communication type is, is also that of their parent. In other words, a literal learner assumes her input has been produced by a literal speaker, and a pragmatic learner assumes her input has been produced by a pragmatic speaker. If biological and cultural evolution were not decoupled, this would mean learners' assumption about their parent's communication type is always right, which makes it easier for them to correctly infer their parent's lexicon and perspective than if they had the wrong assumption about their parent's communication type. Decoupling biological and cultural inheritance therefore stacks the deck against the pragmatic allele that enters the population even more than it already is by the difference in numbers (1 pragmatic agent against 99 literal ones). Learners in the first couple of generations after the pragmatic mutant is introduced would have to be quite lucky to receive not just the pragmatic allele, but *also* input from a pragmatic speaker. This coincidence is more likely to happen the higher the relative fitness of the pragmatic agent(s), and the higher the number of pragmatic agents in the population.

Thus, relative fitness alone will no longer straightforwardly predict the evolutionary stability (and, inversely, the invasibility) of a given agent type in the way that was defined by Maynard Smith and Price (1973). Therefore the current chapter instead uses simulations to asses under what circumstances pragmatic agents might have an evolutionary advantage over literal agents. Specifically, we can run large batches of independent simulation runs in which a pragmatic mutant is introduced in a population of literal agents, and looks at whether the pragmatic allele fixates in more of those populations than would be expected by genetic drift alone. If the number of populations in which the pragmatic allele fixates is lower than would be expected by genetic drift, we can infer that literal agents have a fitness advantage over pragmatic agents. If the number of populations in which the pragmatic allele fixates is instead higher than expected by genetic drift, we can infer that pragmatic agents have a fitness advantage over literal agents, and are therefore 'evolvable' in the corresponding condition. Finally, if the number of populations in which the pragmatic allele fixates is not markedly different from what would be expected by drift, communication type is probably selectively neutral.

Genetic drift is the baseline to which we compare, because this is what happens when the average fitness of an individual is not affected by which allele it carries, and thus changes in allele frequency will be random. The probability that a new allele entering the population will reach fixation if it is selectively neutral (i.e. by drift alone) is equal to the allele's initial frequency when entering the population (Barton et al., 2007, chapter 18, p. 493).[1] For a batch of independent populations, the probability that the mutant allele fixates in a given number of populations is then given by the probability mass function of the binomial distribution as shown in Equation 6.1.

$$P(k; n, p) = \binom{n}{k} p^k (1-p)^{n-k} \tag{6.1}$$

where p is the probability of fixation of the mutant allele in a single population, n is the number of independent populations (i.e. simulation runs), and k is the number of populations in which the mutant allele reaches fixation (i.e. the number of 'successes').

6.3 Pragmatic agents have an evolutionary advantage under both selection for communication and selection on perspective-inference

To measure how likely it is that pragmatic reasoning evolves in a population of literal agents who have already converged on a set of lexicons, we performed an invasibility analysis. As discussed in Chapter 4, convergence is defined here as a state of the population in which fluctuations in the average informativeness of its lexicons remains within a range of 0.1 *ca* for at least 50 consecutive generations. After this fixation point was reached, a single pragmatic agent was then introduced in the populations (such that the make-up of the population changed from 100 literal agents to 1 pragmatic agent and 99 literal agents), and the simulation was subsequently continued in its corresponding

[1] The fact that the probability of fixation of an allele under genetic drift is equal to its initial frequency can be illustrated with a genealogical example. Imagine a population of size N of haploid individuals who each have either of two alleles (i.e. genetic variants) a or b. Let N_a be the number of individuals with allele a and N_b the number of individuals with allele b. When this population reaches fixation on one of the two alleles, we know that all agents in the population at time t_{fix} are descendants from either only t_0 agents with allele a, or only t_0 agents with allele b. If the alleles are selectively neutral (i.e. if their spread through the population is characterised by genetic drift only), the probability of going to fixation is equal for each individual allele present in generation t_0 (i.e. $P_{fix}(a_1) = P_{fix}(a_2) = ... = P_{fix}(a_{N_a}) = P_{fix}(b_1) = P_{fix}(b_2) = ... = P_{fix}(a_{N_b})$). In other words, $P_{fix}(a_i) = P_{fix}(b_i) = \frac{1}{N_a+N_b}$, from which it follows that the probability of allele a fixating in the population is equal to its initial frequency: $P_{fix}(a) = \frac{N_a}{N_a+N_b}$. And the same holds for allele b: $P_{fix}(b) = \frac{N_b}{N_a+N_b}$.

selection condition for another 200 generations. As shown below, this was a sufficient amount of generations in order for one of the two alleles to reach fixation in 598 out of 600 populations.

Table 6.1 shows how many out of 200 independent simulation runs ended up with the pragmatic allele reaching fixation, together with the probability of that happening under the assumption of genetic drift (as given by the probability mass function of the binomial distribution, shown in Equation 6.1). This table shows that the probability that number of populations in which the pragmatic allele reached fixation was produced by drift is very low in the *Selection for communication* condition, and extremely low in the *Selection on perspective-inference condition.*

Table 6.1: k = number of populations in which pragmatic allele reaches fixation, out of the total number of runs that go to fixation in either direction (which is 198 out of 200 runs in the case of the *No Selection* condition, and the full 200 runs in the case of both selection condition). p_{drift} = probability of pragmatic allele fixating in k populations under the assumption of pragmatic allele being selectively neutral, i.e. under genetic drift.

Number of populations in which pragmatic allele fixates					
No Selection		Selection for communication		Selection on perspective-inference	
k=0 (of 198)	p_{drift}=$1.367e^{-01}$	k=9 (of 200)	p_{drift}=$1.724e^{-04}$	k=15 (of 200)	p_{drift}=$2.279e^{-09}$

Table 6.2 shows the 95% confidence intervals for the probability of the pragmatic allele reaching fixation, based on the results summarised in Table 6.1. The confidence intervals for the two selection conditions do not overlap with those for the *No selection* condition, which indicates that the underlying probability of the pragmatic allele fixating in the population is very likely to indeed be higher in the two selection conditions than it is in the *No selection* condition (i.e. under drift).

Table 6.2: 95% confidence intervals for the probability of fixation of the pragmatic allele in the different conditions, based on observed fixation frequencies as shown in Table 6.1. These confidence intervals were obtained using the Poisson approximation for the binomial distribution, because $n \cdot p_{drift} < 5$ (where n is the number of independent simulation runs per condition, and p_{drift} is the probability of the pragmatic allele fixating under drift). (Note that the confidence interval obtained in the *No selection* condition includes the probability of fixation under drift: $1/100 = 0.01$.)

95% confidence intervals for fixation frequencies of pragmatic allele					
No Selection		Selection for communication		Selection on perspective-inference	
lower limit	upper limit	lower limit	upper limit	lower limit	upper limit
0.0000	0.0185	0.0206	0.0854	0.0420	0.1237

Figure 6.1 shows the timecourses of how the frequency of the pragmatic allele changes over time after the single pragmatic mutant has been introduced. This figure shows first of all that in the absence of any selection pressure, the pragmatic allele can reach frequencies as high as 0.8 without that leading to the allele fully taking over (at least not within the timeframe of 200 generations). Furthermore, in 2 out of 200 simulation runs the two different alleles continue to coexist, with fluctuating relative frequencies, without either of them fixating within the timeframe of 200 generations. In other words, the *No selection* condition bears all the marks of genetic drift, as would be expected. After all, both biological and cultural parents are selected at random in this condition, rather than on the basis of their communicative or perspective-inference success. Under the two selection conditions in contrast, all simulation runs in which the frequency of the pragmatic allele exceeds a proportion of 0.25, this allele ends up reaching fixation. Thus, once a quarter of the population shares the pragmatic allele, being pragmatic carries a definite evolutionary advantage under both a selection pressure for communication and a selection pressure on perspective-inference. Figure 6.1 further shows that it takes an average of about 80 generations for the pragmatic allele to fully take over and push out the literal allele. This progression again looks similar across the two selection conditions.

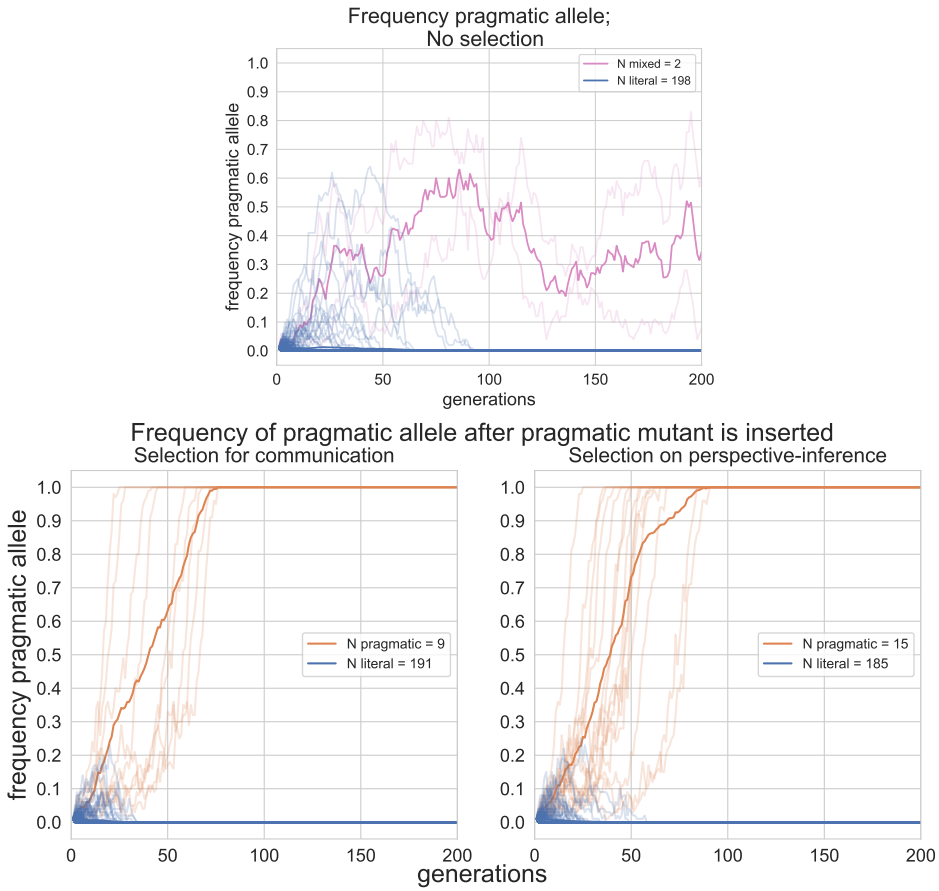


Figure 6.1: Proportion of pragmatic agents in population (i.e. frequency of pragmatic allele) over generations after pragmatic mutant is inserted. Thus, generation 0 in these plots is the generation at which the pragmatic mutant is inserted, but this has been preceded by the number of generations necessary for each of the literal populations to reach convergence in terms of the average informativeness of their lexicons. Subplots show 200 independent simulation runs each, coloured by which allele eventually fixates in the population: literal, pragmatic, or neither. Dark coloured lines show mean for each subgroup, and light coloured lines show the individual simulation runs within that subgroup.

Figure 6.2 shows how the average informativeness of the lexicons in the population develops after the pragmatic mutant has been inserted. In the *No selection* condition, there is no difference in average informativeness between the populations in which the literal allele fixates, and those populations in which the literal and pragmatic allele continue to coexist. In the two selection conditions in contrast, a clear difference in average informativeness emerges between those populations in which the literal allele fixates and those in which the pragmatic allele fixates. In the latter population type, the average informativeness of the lexicons drops from the level that is normally converged on by literal populations under the respective selection pressure as shown in Chapter 4, to the level that is normally converged on by pragmatic populations as shown in

Chapter 5. In other words, populations in which the pragmatic allele takes over do not maintain set of highly informative lexicons that they start out with. Instead, due to the fact that pragmatic populations can make do with less informative lexicons, as shown in Chapter 5, selection on informative lexicons becomes less strict as the proportion of pragmatic agents in the population increases.

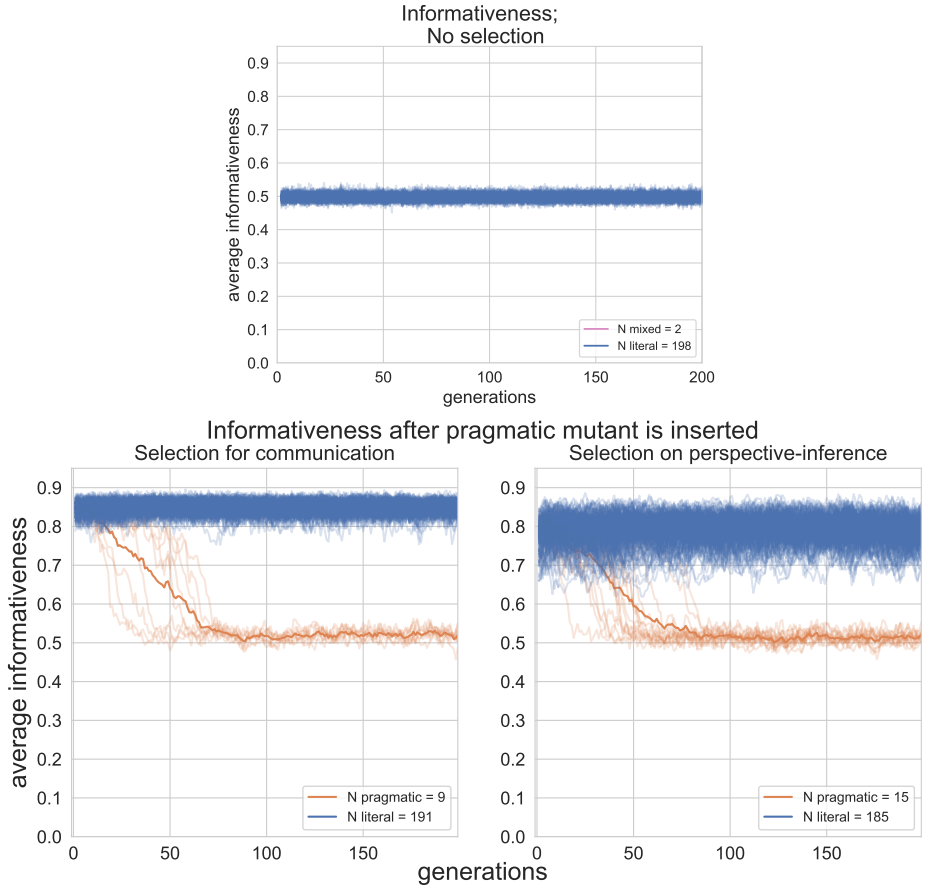


Figure 6.2: Average informativeness of lexicons in population over generations after pragmatic mutant is inserted. Thus, generation 0 in these plots is the generation at which the pragmatic mutant is inserted, but this has been preceded by the number of generations necessary for each of the literal populations to reach convergence in terms of the average informativeness of their lexicons. Subplots show 200 independent simulation runs each, coloured by which allele eventually fixates in the population: literal, pragmatic, or neither. Dark coloured lines show mean for each subgroup, and light coloured lines show the individual simulation runs within that subgroup.

The sample of lexicons that the evolved-to-be-pragmatic populations end up with is not random however. Figure 6.3 shows the extent to which the lexicons in the different populations exploit the benefit in terms of communicative success that comes with being a pragmatic communicator. This 'pragmatic benefit' of a lexicon (PB_ℓ) is defined as the average difference in communicative success between a pair of pragmatic agents

and a pair of literal agents with that lexicon (averaged over whether the listener has the correct or incorrect model of the speaker's perspective), as shown in Equation 6.2.

$$PB_\ell = \frac{1}{|P|} \sum_{p' \in P} CS(S_1, L_2 \mid \ell, p') - CS(S_0, L_1 \mid \ell, p') \tag{6.2}$$

where p' stands for the listener's model of the speaker's perspective, $|P|$ is the total number of perspective hypotheses that learners consider, and CS stands for communicative success. (See also Figure 5.2 in Chapter 5 for a summary plot of how communicative success differs between each of these possible speaker-listener pairs for each of the possible lexicon types.)

As Figure 6.3 shows, this pragmatic benefit is maximised by populations in which the pragmatic allele fixates. This effect is strongest in the *Selection for communication* condition, which makes sense given that this is the condition that directly selects agents on their communicative success. A slight divergence in pragmatic benefit between the literal and pragmatic populations is visible in the *Selection on perspective-inference* condition as well however, which is explained by the fact that the same lexicon types which maximise communicative success for pragmatic agents, also maximises their ability to learn about their parent's perspective, as shown in Chapter 5.

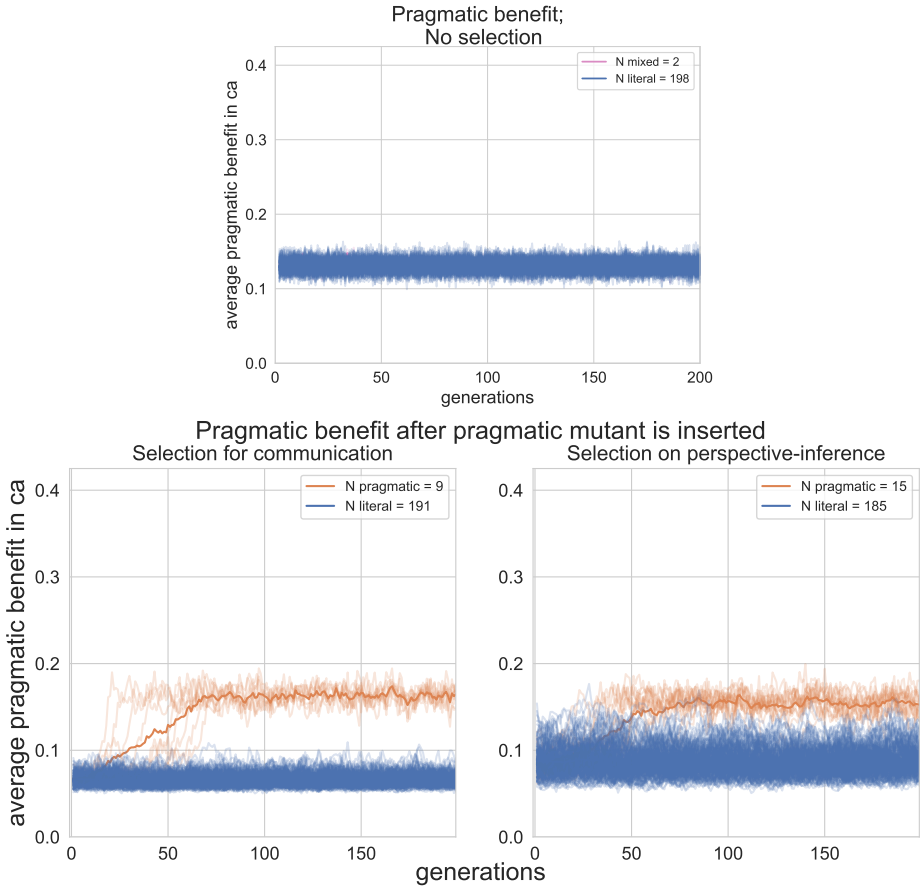


Figure 6.3: Average pragmatic benefit of lexicons in population (i.e. extent to which being used by pragmatic communicators increases the communicative success that can be reached with the lexicon relative to literal communicators) over generations after pragmatic mutant is inserted. Thus, generation 0 in these plots is the generation at which the pragmatic mutant is inserted, but this has been preceded by the number of generations necessary for each of the literal populations to reach convergence in terms of the average informativeness of their lexicons. Subplots show 200 independent simulation runs each, coloured by which allele eventually fixates in the population: literal, pragmatic, or neither. Dark coloured lines show mean for each subgroup, and light coloured lines show the individual simulation runs within that subgroup.

Finally, Figure 6.4 shows how the changes in the selection of lexicons that happen in the populations in which the pragmatic allele fixates affect their fitness. This figure reveals that although these populations select their lexicons in a way that maximises the benefits of being pragmatic, their average fitness (in the sense of the success score that is used to determine agents' probability of becoming a cultural parent) becomes markedly lower than that of the literal populations. This is in line with the results of pragmatic populations' average success at communicating and inferring perspectives as reported in Chapter 5: although pragmatic agents can make do with certain more ambiguous lexicon types, their average success does not reach the same levels as that

of literal populations, at least when the bottleneck width for both population types is kept constant.

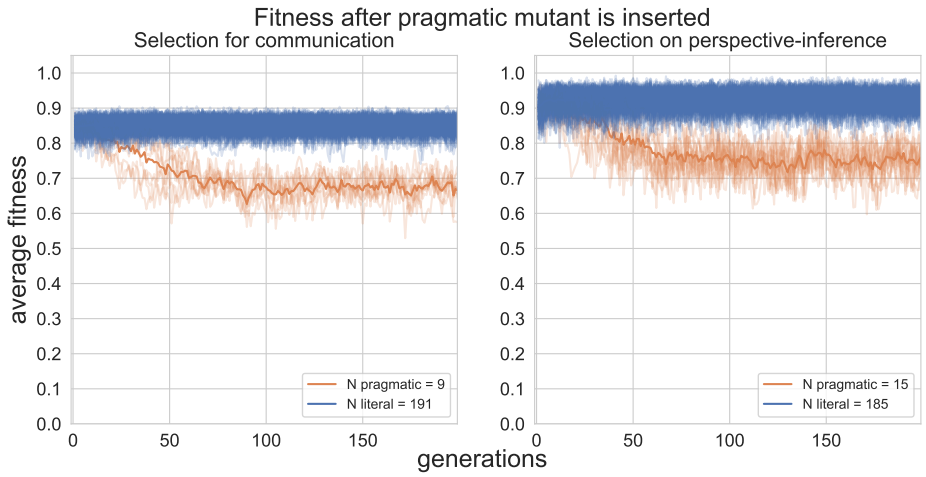


Figure 6.4: Average fitness of the population over generations after pragmatic mutant is inserted. Thus, generation 0 in these plots is the generation at which the pragmatic mutant is inserted, but this has been preceded by the number of generations necessary for each of the literal populations to reach convergence in terms of the average informativeness of their lexicons. Subplots show 200 independent simulation runs each, coloured by which allele eventually fixates in the population: literal, pragmatic, or neither. Dark coloured lines show mean for each subgroup, and light coloured lines show the individual simulation runs within that subgroup. The *No selection* condition is not shown because fitness in this condition is undefined.

6.4 Discussion

The current chapter explored the question of the 'evolvability' of pragmatic agents: Under what circumstances do pragmatic communicators have an evolutionary advantage over literal communicators that would be big enough for them to 'take over'? In order to answer this question, this chapter used a model of gene-culture co-evolution, where pragmatic ability is genetically inherited while lexicons are transmitted culturally. The results obtained with this model showed that whereas being pragmatic is selectively neutral in the absence of any selection pressure, it carries an evolutionary advantage both under a pressure for successful communication and under a pressure for correct perspective-inference. This evolutionary advantage is revealed by the fact that if a pragmatic 'mutant' is inserted in a population of literal agents who have reached convergence on a set of lexicons, the genetic variant of the pragmatic mutant ends up spreading through the population and ultimately taking over more often than would be expected by genetic drift alone (i.e. if being pragmatic carried no evolutionary

advantage).

As this pragmatic allele spreads through the population, the selection of lexicons in the population changes: more ambiguous lexicons are adopted. Exactly which of these more ambiguous lexicons are adopted is not random however: populations with a majority of pragmatic agents select their lexicons in such a way that the benefits that come with being pragmatic (in terms of communicative success and accumulating information about others' perspectives) is maximised. However, this maximisation of pragmatic benefit does not mean that the agents in populations which have evolved to be entirely pragmatic reach higher or even equal levels of fitness as agents in literal populations. Instead, pragmatic populations' average fitness ends up being lower than that of literal populations. Thus, the evolutionary advantage of being pragmatic in the current model with the current parameter settings is not an *absolute* advantage. Rather, it is an advantage that exists in certain circumstances, but can subsequently lock populations in in a suboptimal state.

In the simulations reported above, the first pragmatic mutant enters a population of literal agents who have converged on a selection of highly informative lexicons. As shown in Chapter 4, the stationary distribution of lexicon types in literal populations after convergence is very strongly skewed in the direction of the most informative lexicon type, under both selection conditions. The latter is a lexicon type for which pragmatic agents only have a slight advantage over literal agents in terms of communicative success and perspective-inference. This slight advantage can cause the pragmatic allele to start spreading through the population however, and as it spreads, the pressure on maintaining this maximally informative lexicon type relaxes (because pragmatic agents can reach decent levels of fitness even when receiving input from a less informative lexicon). This causes the proportion of less informative lexicon types to spread through the population, which further increases pragmatic agents' relative fitness compared to literal agents, and so on. Thus, once the pragmatic allele starts spreading through the population, the population enters a positive feedback loop in which the number of pragmatic agents in the population and the number of ambiguous lexicons in the population promote each other. The simulation results reported above indicate that once populations exceed a threshold of a quarter of the agents in the population being pragmatic, there is no turning back, even though the overall fitness of the population starts decreasing relative to its starting point. A similar effect was described by Lachlan and Slater (1999) in a model of gene-culture co-evolution of an innate ability for vocal

learning and the culturally transmitted songs in songbirds. Lachlan and Slater coined this effect the 'cultural trap' hypothesis: a particular genetically inherited trait (vocal learning in this case) *"is maintained in an evolutionary trap formed by the interaction between genes and culture"* (Lachlan and Slater, 1999, p. 702).

As discussed at the beginning of this chapter, Piantadosi et al. (2012) argue that ambiguity is a desirable feature of a communication system if listeners can make use of the context to make inferences about the speaker's communicative intentions. Piantadosi et al.'s argument is based on considerations of efficiency and ease: a more ambiguous system avoids redundancy and allows for the reuse of elements that are less effortful to produce and comprehend. The current model did not incorporate any explicit pressures in favour of efficiency or ease, but this could be done in future work. Adding a pay-off for using fewer signals, or a prior bias in favour of simpler lexicons, would presumably boost the advantage of the pragmatic allele even further.

The current model did not explicitly add a cost of being pragmatic, even though it would be a very reasonable assumption to make that pragmatic reasoning is costly in terms of the cognitive machinery or processing that is required. However, there is an *implicit* cost to being a pragmatic agent in this model, because pragmatic agents require more observations to correctly infer the lexicon of their cultural parent than literal agents do, as shown in Chapter 5. Pragmatic agents need about twice as many observations to reach the same level of posterior belief in the correct composite hypothesis about their cultural parent as literal agents do. This hurts pragmatic agents in their ability to correctly infer their parent's perspective (although not always; there is in fact a limited set of lexicon types for which pragmatic agents need fewer observations to reach the same level of belief in the correct *perspective* hypothesis as literal agents do, as discussed in Chapter 5). It also hurts pragmatic agents in their ability to communicate with their cultural parent, because slower learning in combination with a transmission bottleneck means lexicons are less likely to be transmitted faithfully between pragmatic agents, leading to pragmatic learners ending up with a lexicon that is different from their parent's more often.

Another aspect that is missing from the current model is the possibility of cultural transmission of pragmatic reasoning. That is, one could envision a model where learners do not only have to infer the lexicon and perspective of their cultural parent, but also whether their cultural parent is a literal or pragmatic speaker. For instance, Brochhagen et al. (2018) used such cultural transmission of 'pragmaticness' in their modelling work

on the co-evolution of lexical meaning and pragmatic use, as described in Chapter 5 (Section 5.1.5). Combining such cultural transmission of pragmatic ability with the current model would allow us to explore how 'evolvable' pragmatic ability is when it is purely culturally transmitted, as hypothesised by Heyes and Frith (2014) and discussed in more detail in Chapter 2. This would therefore be an interesting avenue for future work with the model presented in this thesis.

Chapter 7

Conclusion

In this thesis I used computational modelling to explore the hypothesis that language and mindreading have co-evolved. This hypothesis has been put forward by several theorists of human evolution on the basis that (i) language use requires an ability to entertain and recognise communicative intentions (and therefore mindreading), to an extent that has (as yet) not been demonstrated in nonhuman animals, and (ii) mindreading benefits from language because language provides us with a wealth of data about what's going on in the minds of others, and with a tool for transmitting our understanding of others' minds to younger members of our population. The aim of this thesis was to formalise the preliminaries of this hypothesised co-evolution in an agent-based model, in order to explore under what circumstances such a co-evolution could have gotten off the ground.

For this purpose, I first presented (in Chapter 3) a new model of word learning in which learners cannot directly observe the speaker's referential intention, nor infer the referent of a novel word through cross-situational learning. Instead, the only way in which learners can accurately infer the mappings from referents to signals is by learning about the speaker's perspective on the world. This perspective, in combination with the context, generates a private, subjective 'mental state' in the speaker in the form of a probability distribution over potential referents. (This is a distribution in which every referent that exists in the agents' world has a nonzero probability of being the speaker's intended referent, in every context.) A speaker's perspective, however, is a hidden, unobservable variable. The only data the learner receives that could help them infer that perspective consists of the speaker's utterances in context. Therefore, if the learner knows what lexicon the speaker is using, it would be fairly easy for them to infer

the speaker's perspective (given that they can observe the context in each interaction). And vice versa, if the learner knows the speaker's perspective, it would be fairly easy for them to infer the speaker's lexicon. However, learners in this model have to infer both attributes of the speaker simultaneously. The simulation results presented in Chapter 3 show that Bayesian learners can solve this joint-inference task (given enough data from different contexts), but only if two conditions are met. Firstly, the learner has to be able to represent the speaker's perspective, and secondly, the speaker's lexicon has to be at least somewhat informative. Thus, lexicon-learning and perspective-learning co-develop in this model: one cannot happen without the other.

These results led to a question about language emergence: if this co-development depends on learners receiving input from a lexicon that contains mappings that are somewhat informative (i.e. not entirely ambiguous), how could a population of such learners evolve an informative lexicon from scratch? This question was explored in Chapter 4 by embedding the developmental model described above in a model of iterated learning — where lexicons are passed on over generations through observational learning. Simulation results obtained with this model showed that if there is no pressure for populations to evolve an informative lexicon, they will end up with a random sample from all possible lexicons. However, if populations are exposed to a selection pressure for *either* successful communication *or* successful perspective-inference, they converge on lexicons of the most informative type (i.e. lexicons with only one-to-one mappings between referents and signals).[1] In both cases, this evolution of informative lexicons leads not just to improvement of the skill that is selected for (i.e. communication or perspective-inference respectively), but also to improvement of the remaining skill. This is a consequence of the fact that both successful communication and successful perspective-inference rely on learners receiving input from an informative lexicon. In the case of communication, more informative lexicons provide listeners with more information about a speaker's intended referent, and also make it more likely that learners correctly infer the lexicon of their cultural parent (given a bottleneck on transmission). In the case of perspective-inference, more informative lexicons allow learners to accumulate more information about a speaker's perspective.

[1] Agents' fitness under both selection pressures is measured on the basis of the agent in relation to their cultural parent. In the case of selection for communication, agents' fitness is determined by how successful they are at interpreting their cultural parent's utterances. In the case of selection on perspective-inference, agents' fitness is determined by how much posterior belief they assign to the correct hypothesis about their cultural parent's perspective.

In Chapter 5, the model of communication was extended to include pragmatic reasoning. This allowed speakers to optimise their utterance choice on the basis that listeners will take their perspective when interpreting those utterances. In turn, pragmatic listeners reason about such a pragmatic speaker when interpreting the speaker's utterances. Pairs of pragmatic communicators can reach higher levels of communicative success compared to pairs of literal communicators (as were used in Chapters 3 and 4), because their pragmatic reasoning can compensate for ambiguity in the lexicon. This causes populations of pragmatic agents to converge on more varied sets of lexicons under the two selection pressures than literal populations do, while still reaching decent levels of success at communicating and inferring perspectives. The lexicons that evolve in these pragmatic populations look more like the natural languages we find today, in the sense that they contain ambiguity which can be resolved if listeners take into account the context and the speaker's point of view, and speakers know and rely on this. As argued by Piantadosi et al. (2012), establishing a less ambiguous language under such circumstances would in fact be costly.

Finally, Chapter 6 explored under what circumstances these pragmatic agents could have an evolutionary advantage over literal agents, using a model of gene-culture coevolution and an invasibility analysis. Simulation results obtained with this model showed that pragmatic agents have an evolutionary advantage under both a selection pressure for communication and a selection pressure on perspective-inference: In the former case because pragmatic agents can reach higher levels of communicative success given a particular lexicon than literal agents can, and in the latter case because given certain lexicons, pragmatic speakers give away more information about their own perspective in their utterance productions than literal speakers do. However, as the number of pragmatic agents in a population grows relative to the number of literal agents, the populations' overall success at communicating and inferring each others' perspectives does not increase. Instead, the populations' average success on both measures drops somewhat. This is a result of populations entering a 'cultural trap' (Lachlan and Slater, 1999): as the number of pragmatic agents in the population increases, the pressure on lexicons to be unambiguous goes down, which in turn results in pragmatic agents having a stronger advantage over literal agents (because pragmatic agents can cope with ambiguous lexicons better than literal agents can). Thus, once pragmatic reasoning in communication becomes a possibility, the lexicon adapts in such a way that pragmatic reasoning subsequently becomes a necessity.

The simulation results summarised above suggest two potential positive feedback loops between language and mindreading that could drive co-evolution between the two. Firstly, the iterated learning results obtained with literal populations show that, if we assume that linguistic input provides agents with data that helps them learn about others' point of view, and that understanding of others' point of view in turn is important for language-learning, the emergence of an informative lexicon will not only improve populations' communication but also their perspective-inference. In turn, improved perspective-inference causes more successful communication and more faithful lexicon transmission. This first positive feedback loop is a result of cultural evolution alone: nothing changes in the agents' underlying ability to learn about perspectives (i.e. their genetic endowment). What changes instead is the observational data that agents receive: observing the utterances-in-context of speakers who use a more informative lexicon provides learners with more information about those speakers' perspective.

The second positive feedback loop is between pragmatic reasoning and language, as shown by the invasibility results obtained with the gene-culture co-evolution model. As the pragmatic reasoning skills of a population increase, cultural evolution will cause the populations' lexicons to evolve in a way that maximises the benefits of such pragmatic reasoning (which, in the case of a compositional language that is used to express a potentially infinite amount of different possible meanings, would increase its ease and efficiency, as argued by Piantadosi et al., 2012), consequently making pragmatic reasoning more and more indispensable. This second positive feedback loop is — as it arises in the modelling work presented in this thesis — a result of the combination of biological and cultural evolution. That is, populations' pragmatic reasoning abilities increase as a result of biological evolution (due to a pragmatic 'allele' spreading through the population through genetic inheritance), while populations' lexicons change as a result of cultural evolution. However, the biological aspect of this dynamic is not a requirement: even if humans' pragmatic reasoning abilities are culturally transmitted, as argued by Heyes and Frith (2014), the same positive feedback loop could still ensue.

In sum, co-evolution between lexicons and perspective-inference arises in the model presented in this thesis under both selection for communication and selection on perspective-inference. Either such pressure, and potentially both, could have arisen during the Pleistocene epoch when our ancestors of the *Homo* lineage started adopting a more interdependent lifestyle (with collaborative foraging as its prime example), which required increasingly sophisticated coordination and communication (Sterelny, 2012; Tomasello

et al., 2012; Whiten and Erdal, 2012). The two selection pressures are different, however, in terms of whether they *assume* coordination. Mindreading is useful in any population in which individuals interact, no matter whether these interactions are coordinated, because it allows individuals to predict (and possibly manipulate) others' behaviour and use this to their own benefit. This is different for language: a conventional communication system such as language only pays off once it is shared with others; i.e. once some threshold level of coordination has been reached (Sterelny, 2012). Selection on mindreading skills as required for coordination and intentional (but not conventional) communication is thus arguably a more plausible starting point than selection on language (i.e. conventional communication).

However, the simulation results presented in this thesis suggest that even just selection on one of these two skills could kick off a positive feedback loop which results in further sophistication of both. Such a two-way positive feedback relationship between language and mindreading is in line with the theoretical scenarios presented by Sterelny (2012), Tomasello et al. (2012) and Whiten and Erdal (2012). The two selection pressures as implemented in this model can be interpreted as either biological selection (where more successful agents are more likely to have offspring, and those offspring learn their lexicon from their biological parents) or cultural selection (where new agents who enter the population choose which agent of the previous generation they want to receive their data from, and more successful agents are more likely to be chosen as such cultural parents).

The power of computational modelling lies in part in simplifying reality in such a way that it allows us to get a grip on complex systems. As a result of this simplification, there are several things missing from the modelling work presented in this thesis, some of which are promising avenues for future research. Firstly, agents learn about their cultural parent's perspective and lexicon through purely observational learning, not interaction. This is a design feature of iterated learning models with Bayesian agents, on which there is a lot of prior work that this thesis builds on (as reviewed in Chapter 4). However, it is not a very realistic assumption. In real life, language users receive constant feedback from each other, which allows learners to check their understanding, and their caregivers to tailor their input to the learner (Yurovsky, 2017). Although this type of feedback is missing from the model used in this thesis, this does not prevent the Bayesian learners from perfectly learning their cultural parent's perspective and lexicon, given that they receive enough data (except for learning the parent's perspective if that

parent is using a completely ambiguous lexicon).

Secondly, in all simulations used in this thesis, learners receive input from only a single speaker or cultural parent. If learners were instead to receive data from multiple agents, it would be reasonable to assume that what they learn about the lexicon and perspective of one such agent would inform them in learning about the lexicon and perspective of another. This could be implemented by extending the current model of learning into a hierarchical Bayesian model (Kemp et al., 2007; Xu et al., 2009). In such a hierarchical model, learners could infer a probability distribution ('overhypothesis') over, for example, possible perspectives in their population. This would mean that with each agent a learner receives data from, the learner would not only update their posterior probability distribution over perspectives for that specific agent, but also a probability distribution over perspectives for the population as a whole. Through such hierarchical updating, a learner who starts out with an egocentric bias could overcome this bias through learning, such that learning about perspectives that are different from the learner's own becomes easier over the course of the learner's lifetime.

Thirdly and finally, as mentioned above, the possibility of cultural transmission of pragmatic reasoning abilities has not been explored in this thesis. It would however be possible to implement with a simple extension of the current model, where learners infer not only the lexicon and perspective of their cultural parent, but also whether their cultural parent produces their signals literally or pragmatically. This would allow for further exploration of the hypothesis that the explicit mindreading skills we find in humans today are the result of cultural, rather than biological, evolution (Heyes and Frith, 2014; Heyes, 2018).

To conclude, language and mindreading are interrelated skills that may well have co-evolved, and this co-evolution may have been driven largely by cultural rather than biological evolution. Together, these two skills also strengthen the potential for further cultural evolution: mindreading is an important aspect of teaching and thus expands the potential for cultural transmission (Heyes, 2018; Dunstone and Caldwell, 2018), and language provides a powerful tool for sharing information, not just within, but also between generations (Whiten and Erdal, 2012). These capabilities underpin the vast amount of cumulative culture we see in humans today: our capacity to pass on knowledge and techniques over generations and refine them in the process. The significance of this capacity I could not phrase better than Heyes (2012a): "Cumulative cultural evolution is what 'makes us odd'" (p. 2181).

Appendix A

Maximally informative contexts

As discussed in Chapter 3 (Section 3.4.3), the way in which learners update their beliefs about speakers' perspectives is by comparing the relative frequencies with which the speaker chooses to talk about the different referents, with those relative frequencies as predicted by the two different perspective hypotheses, over different contexts. This task is further complicated by the fact that the learner does not have direct access to the speaker's intended referent, and instead has to reconstruct this from the speaker's utterances in context, which means that the learner is really comparing the relative frequencies with which the speaker produces different utterances with those relative frequencies as predicted by the different composite (perspective+lexicon) hypotheses. In Bayesian terms, the learner updates its posterior belief in the different perspective hypotheses proportional to the likelihood of observed utterance+context combinations under the different composite hypotheses. This means that the bigger the difference in data likelihood between the two different perspective hypotheses, the quicker the learner can update its posterior belief in the right direction.

The role of the context in creating the differences in data likelihood is that the context is what determines the 'saliency distributions' for the different perspectives. This saliency distribution is what determines the probabilities with which a speaker will choose each object as its intended referent. Because the learner starts out without knowing which utterance maps to which referent, what matters is not how the probabilities of each of the referents differ between the two perspectives, but instead how the relative *ratios* between referent probabilities differ between the two perspectives. Context informativeness is therefore defined as the sum of the absolute difference each such ratio between perspective p_i and perspective p_j as shown in Equation A.1.

$$d(s_{p_i}, s_{p_j}) = \sum_{\mathcal{R}_i \in \rho} |\mathcal{R}_i(p_i) - \mathcal{R}_i(p_j)| \tag{A.1}$$

where each individual ratio $R_i(p_i)$ is given by:

$$\mathcal{R} = P(o_m = r|p_i, c) : P(o_{n \neq m} = r|p_i, c) \tag{A.2}$$

The probability of a given object o_m being chosen as the intended referent r given perspective p and context c ($P(o_i = r|p, c)$) was defined in Chapter 3 as Equation 3.1, replicated below as Equation A.3.

$$P(o_i = r|p, c) = \frac{1 - |p - o_{i_c}|}{\sum_{o \in O} 1 - |p - o_c|} \tag{A.3}$$

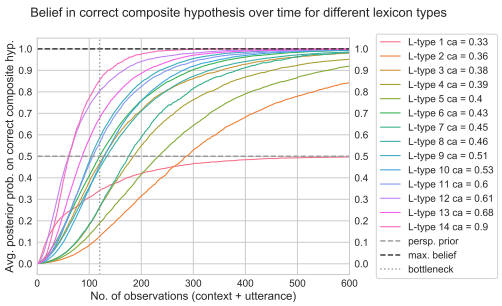
where o_c stands for the attribute of object o in context c (one can think of this attribute as a spatial location), and O stands for the full set of objects. (Recall that all objects are always present in each context.)

The set of maximally informative contexts was chosen by finding the contexts with highest informativeness according to this measure in a finite, discrete search space consisting of all possible contexts created from object attributes in the range $< 0.1, 0.2, 0.3, ..., 0.9 >$ (504 contexts in total, given three objects). This method yielded the maximally informative context $[0.1, 0.2, 0.9]$ (used for the diagram in Figure 3.1 in Chapter 3) and its inverse $[0.1, 0.8, 0.9]$, in all their possible permutations. This resulted in a balanced set of 12 contexts in total, which is simply repeated to produce the desired number of observations for each learner. (With the restriction that the number of observations is always a multiple of 12, such each context occurs with equal frequency.) Although these maximally informative contexts allow languages to be learned faster compared to when contexts are generated randomly (as was the case for the developmental simulations reported in Chapter 3), the learning results for both types of input are similar in qualitative terms (see figures B.1 to B.4 in Appendix A).

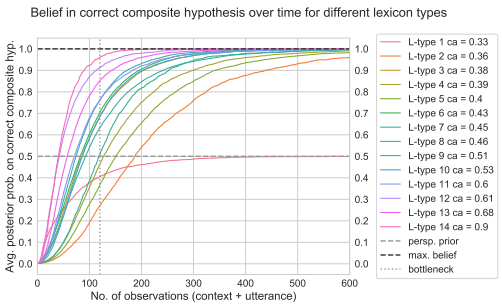
Appendix B

Learning from maximally informative contexts

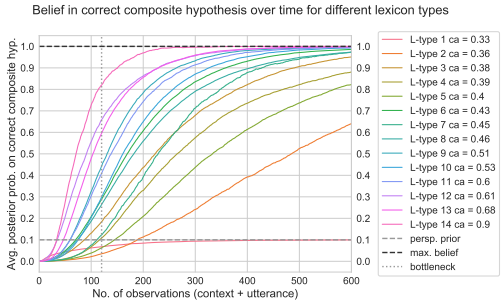
This appendix shows the same learning curves as shown in Chapter 3, but compares the results for learners observing only repetitions of a fixed set of maximally informative contexts (as used for the simulations in the current chapter) with the results for learners observing randomly generated contexts (as used for the simulations reported in Chapter 3). Figures B.1 through B.2 demonstrate that although observing only maximally informative contexts speeds up learning, the qualitative results do not change. Figure B.4 shows the same but for the number of observations required to learn the different lexicon types.



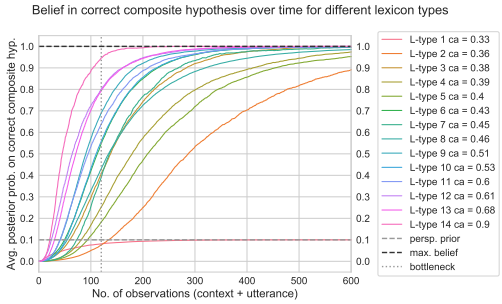
(a) Random contexts; unbiased learner



(b) Optimal contexts; unbiased learner

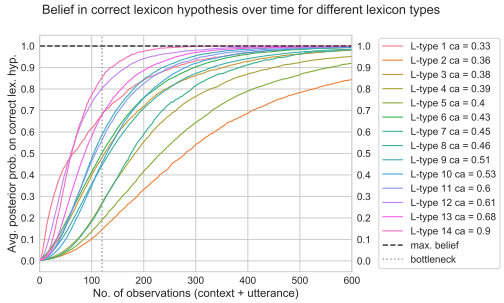


(c) Random contexts; egocentric learner

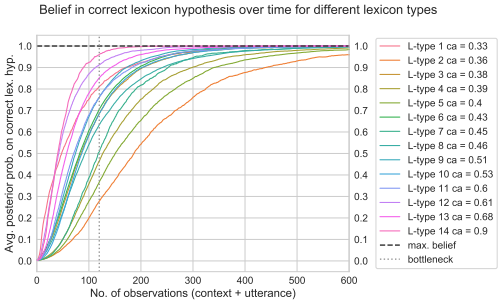


(d) Optimal contexts; egocentric learner

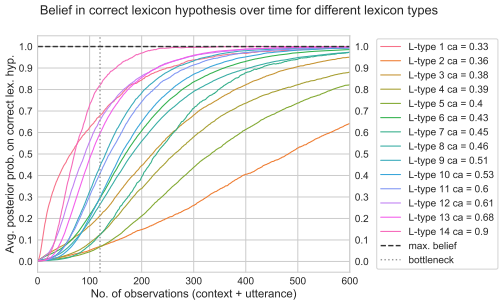
Figure B.1: Learning curves showing posterior probability assigned to the correct composite hypothesis (i.e. lexicon + perspective) over time for learners observing randomly generated contexts (a and c) and learners observing only 'optimal' contexts (b and d). Graphs show learning results for all different possible input lexicons, categorised by informativeness class. *ca* levels of different lexicon types range from lowest possible (0.33333...) to highest possible (0.90375) for lexicon size = 3x3 and error rate $\epsilon = 0.05$. Lines show grand means over all lexicons within a given informativeness class, and 100 independent simulation runs per individual input lexicon. Grey dashed line indicates the prior probability assigned to the correct perspective hypothesis. Black dashed line indicates maximum posterior probability that can be reached.



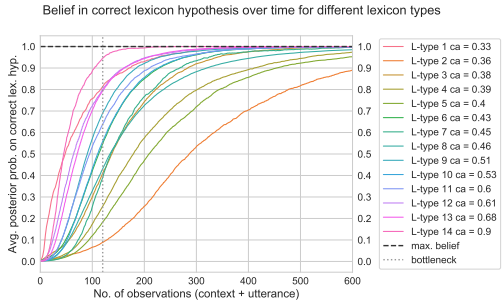
(a) Random contexts; unbiased learner



(b) Optimal contexts; unbiased learner

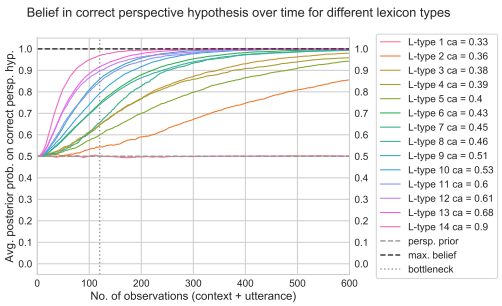


(c) Random contexts; egocentric learner

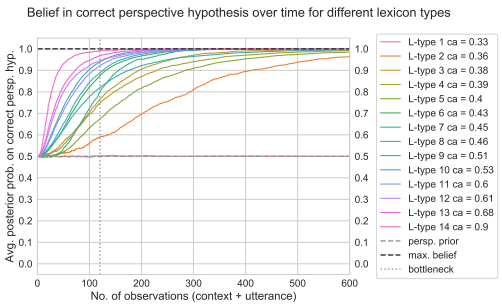


(d) Optimal contexts; egocentric learner

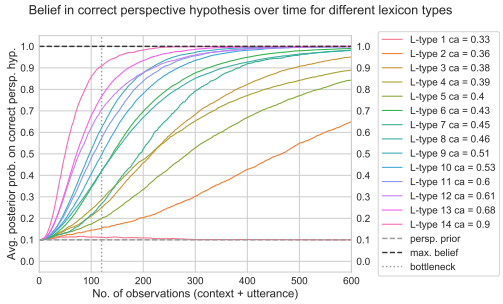
Figure B.2: Learning curves showing posterior probability assigned to the correct lexicon hypothesis over time for learners observing randomly generated contexts (a and c) and learners observing only 'optimal' contexts (b and d). Graphs show learning results for all different possible input lexicons, categorised by informativeness class. *ca* levels of different lexicon types range from lowest possible (0.33333...) to highest possible (0.90375) for lexicon size = 3x3 and error rate $\epsilon = 0.05$. Lines show grand means over all lexicons within a given informativeness class, and 100 independent simulation runs per individual input lexicon.



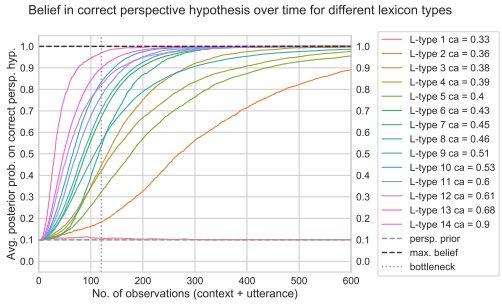
(a) Random contexts; unbiased learner



(b) Optimal contexts; unbiased learner

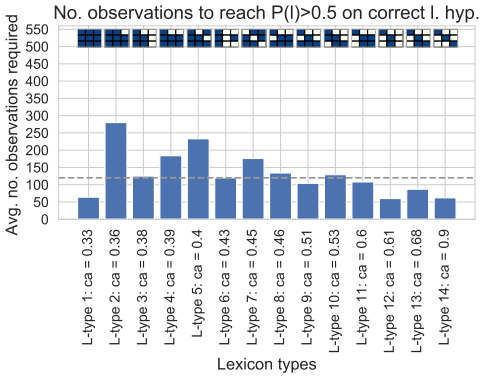


(c) Random contexts; egocentric learner

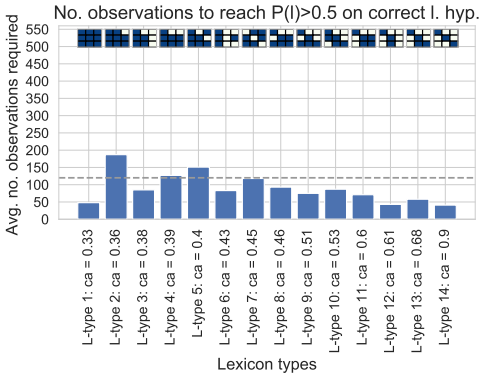


(d) Optimal contexts; egocentric learner

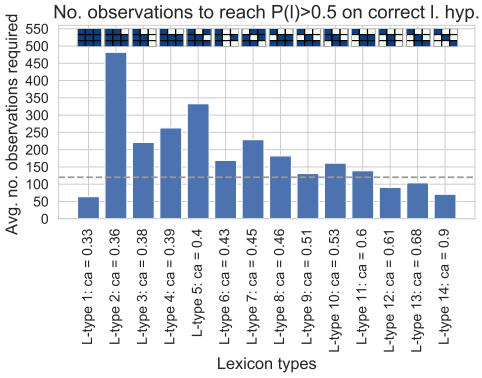
Figure B.3: Learning curves showing posterior probability assigned to the correct perspective hypothesis over time for learners observing randomly generated contexts (a and c) and learners observing only 'optimal' contexts (b and d). Graphs show learning results for all different possible input lexicons, categorised by informativeness class. *ca* levels of different lexicon types range from lowest possible (0.33333...) to highest possible (0.90375) for lexicon size = 3x3 and error rate $\epsilon = 0.05$. Lines show grand means over all lexicons within a given informativeness class, and 100 independent simulation runs per individual input lexicon. Grey dashed line indicates the prior probability assigned to the correct perspective hypothesis. Black dashed line indicates maximum posterior probability that can be reached.



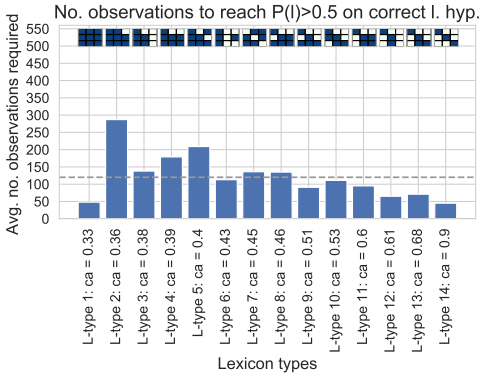
(a) Random contexts; unbiased learner



(b) Optimal contexts; unbiased learner



(c) Random contexts; egocentric learner



(d) Optimal contexts; egocentric learner

Figure B.4: Average number of observations required to reach the threshold of $P(\ell) > 0.5$ posterior belief in the correct lexicon hypothesis, for learners observing randomly generated contexts (a and c) and learners observing only 'optimal' contexts (b and d). Graphs show learning results for all different possible input lexicons, categorised by informativeness class. *ca* levels of different lexicon types range from lowest possible (0.33333...) to highest possible (0.90375) for lexicon size = 3x3 and error rate $\epsilon = 0.05$. Bars show grand means over all lexicons within a given informativeness class, and 100 independent simulation runs per individual input lexicon.

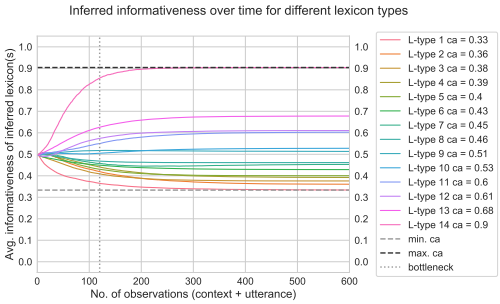
Appendix C

Development of 'inferred informativeness'

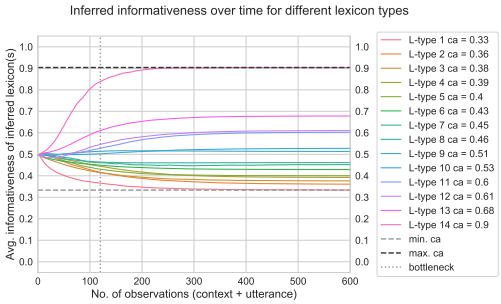
We can also measure how the learner's 'inferred informativeness' develops over time. This is obtained by at each time step multiplying the posterior probability that the learner assigns to each lexicon hypothesis with the informativeness of that lexicon, and summing the resulting value over all lexicon hypotheses, as shown in Equation C.1.

$$ca' = \sum_{\ell \in \mathcal{L}} P(\ell|D) \cdot ca(\ell) \tag{C.1}$$

where ca' stands for the inferred informativeness of ℓ, $\mathcal{L}$ stands for the total space of lexicon hypotheses, $P(\ell|D)$ for the posterior probability of lexicon hypothesis ℓ given data D, and $ca(\ell)$ for the informativeness of lexicon ℓ, measured as described in Chapter 3 (Section 3.5). This measure of inferred informativeness allows us to see how quickly the learner's belief about the informativeness of the input lexicon reflects its actual informativeness, as shown in Figure C.1.



(a) Uniform prior



(b) Egocentric perspective prior

Figure C.1: Average informativeness of 'inferred' lexicon over time, for all different possible input lexicons, categorised by informativeness. *ca* levels of different lexicon types range from lowest possible (0.33333...) to highest possible (0.90375) for lexicon size = 3x3 and error rate $\epsilon = 0.05$. Grey dashed line indicates the minimum informativeness that a lexicon can have (equal to chance level for three referents); black dashed line indicates the maximum informativeness a lexicon can have.

As Figure C.1 shows, the learner's inferred informativeness becomes equal to the actual informativeness of the input lexicon relatively quickly (roughly between 250 and 400 observations), and the variation in the rate at which this happens is much less great, and less dependent on the informativeness of the input lexicon than the variation in learning rates that can be observed in figures B.1 through B.4 in Appendix A.

Appendix D

Confusability of lexicon types: literal speakers

The lexicon types that are overrepresented in populations of literal, egocentric agents in the *No selection* condition, as shown in Figure 4.5, are types 2 and 5 (where overrepresentation is defined as a lexicon type being selected with a proportion of more than one standard deviation above the mean[1]). This is surprising at first sight given that these are the two lexicon types which take the most observations to learn; the mean number of observation required to reach $P(\ell) > 0.5$ is larger than the bottleneck width for each of these lexicon types, for both egocentric and unbiased learners (see figures 4.2 and B.4). Thus, out of all lexicon types, types 2 and 5 are the least likely to be transmitted faithfully, which seems at odds with their being overrepresented.

However, the amount of observations required to learn a lexicon type is not the only factor that predicts how likely agents are to select it in the process of iterated learning. Another important factor is the probability that the lexicon will be transitioned into, given input from another lexicon. If the data produced by one or more lexicon types j is likely to be 'misinferred' as coming from lexicon type i, type i does not need to be faithfully transmitted itself. As mentioned in Chapter 4 (Section 4.3.1), calculating the exact transition matrix for the current model using the equations derived by Griffiths and Kalish (2007), becomes intractable for lexicon sizes larger than 2x2 combined with a large amount of observations. However, we can use just the production probability part of this procedure to derive a measure of the *confusability* between the data of one lexicon with that of another.

[1]For lexicon type 2 this proportion is also bigger than two standard deviations above the mean.

This confusability can be formalised as the inverse of the difference between the data predicted by one lexicon type LT_i and another lexicon type LT_j. That is, the smaller this difference in predicted datasets, the more confusable the two lexicon types are. Given that learners in the simulations reported in Chapters 4, 5 and 6 observe only repetitions of a fixed set of 12 maximally informative contexts, we can simply compute the probability of each utterance given each of these contexts and compare these probabilities between lexicons. The sum of the absolute difference between utterance-in-context probabilities for two different lexicons provides a direct measure of how different the data produced by these two lexicons will be, as shown in Equation D.1. This yields a difference measure $\delta_c(l_i, l_j)$ for lexicon l_i and lexicon l_j given context c. (See Equation 3.2 in Chapter 3 for the likelihood of an utterance s given a lexicon, perspective and context, $P(s \mid l, p, c)$.)

$$\delta_c(\ell_i, \ell_j) = \sum_{s \in \mathcal{S}} |P(s \mid \ell_i, p, c) - P(s \mid \ell_j, p, c)| \tag{D.1}$$

(Here we consider only the perspective p that corresponds to speakers' true perspective in the simulations.) $\delta(\ell_i, \ell_j)$ is obtained by taking the sum of the difference in predicted data $\delta_c(\ell_i, \ell_j)$ over the full set of maximally informative contexts that learners observe (12 in total), as shown in Equation D.2.

$$\delta(\ell_i, \ell_j) = \sum_{c \in C} \delta_c(\ell_i, \ell_j) \tag{D.2}$$

If $\delta(\ell_i, \ell_j)$ is small, the likelihood of the data will be high under both hypothesis ℓ_i and hypothesis ℓ_j for a Bayesian learner (because the likelihood is defined using the actual production algorithm that generates the data). Therefore, when $\delta(\ell_i, \ell_j)$ is small, a large amount of observations will be required in order for the hypothesis that does not correspond to reality to relinquish posterior probability. To group this data difference measure by lexicon type, $\delta(LT_i, LT_j)$ is obtained by taking the sum over each possible combination of lexicons of type LT_i and lexicons of type LT_j (while dividing by the number of lexicons in each type $|LT|$, because this differs per lexicon type), as shown in Equation D.3.

$$\delta(LT_i, LT_j) = \frac{1}{|LT_i|} \frac{1}{|LT_j|} \sum_{\ell_i \in LT_i} \sum_{\ell_j \in LT_j} \delta(\ell_i, \ell_j) \tag{D.3}$$

As mentioned above, the confusability of two lexicon types is equal to the inverse of the difference in the data they produce, $\delta(LT_i, LT_j)$. Thus, a measure of confusability between lexicon types $C(LT_i, LT_j)$ is obtained by taking the inverse of $\delta(LT_i, LT_j)$ as shown in Equation D.4.

$$C(LT_i, LT_j) = \frac{1}{\delta(LT_i, LT_j)} \tag{D.4}$$

Figure D.1 is the resulting confusability matrix, which shows $C(LT_i, LT_j)$ for each possible combination of lexicon types. This matrix reveals that for every single lexicon type, the data it produces is the most confusable with the data produced by lexicon types 2, 5 and 7. The highest confusability scores in the whole matrix are between these three lexicon types. (That is, the data produced by lexicon type 2 is highly confusable with that produced by types 5 and 7, the data produced by type 5 is highly confusable with that produced by types 2 and 7, and so on.) This, in combination with egocentric learners' bias towards inferring less informative lexicons, explains the fact that lexicon types 2 and 5 are overrepresented in populations of egocentric agents in the *No selection* condition. Although lexicon types 2 and 5 are unlikely to be transmitted faithfully, they are very likely to be transitioned into from other lexicon types, and even more likely to transition into each other.

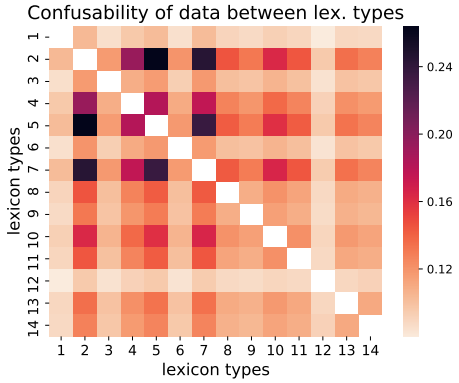


Figure D.1: Confusability matrix showing for each lexicon type in the space of 3x3 lexicons how confusable the data it produces is with the data produced by other lexicon types. Note that the quantity shown in this matrix is not very interpretable given that it is the result of taking the inverse of the sum of the absolute difference between probabilities. We can however interpret the relative differences between the resulting values as relative differences in confusability.

Figure D.1 also sheds light on the distribution over lexicons found in populations of literal agents in the different selection conditions as reported in Chapter 4 (see Figure

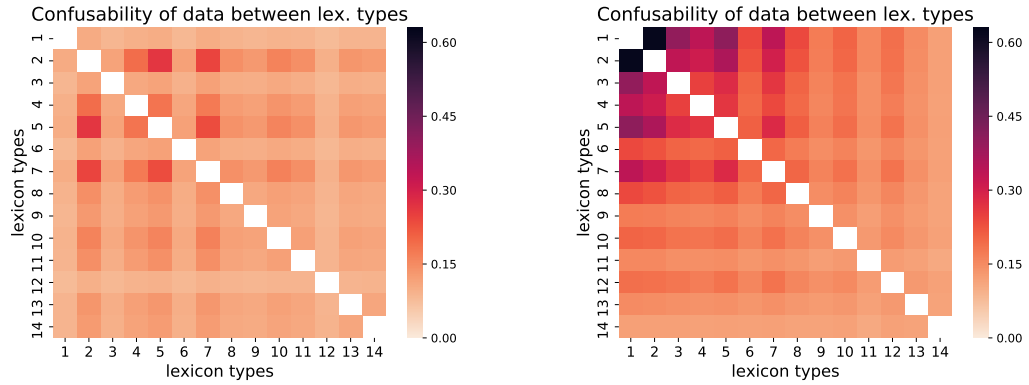
4.5). As mentioned above, the dominant lexicon type in each of these conditions is type 14, but the two lexicon types that most frequently coexist with it are types 7 and 13. The presence of lexicon type 13 ($ca = 0.68$) in these selection conditions can be explained by the fact that this is the next most informative lexicon type after type 14, and will therefore yield relatively high success at lexicon-inference, communication and perspective-inference. On top of that the data it produces is also relatively confusable with that produced by type 14, as shown in Figure D.1. The relatively successful coexistence of lexicon type 7 on the other hand is more surprising at first blush, given that it has a relatively low informativeness level (ca=0.45) and is not particularly easy to learn (see figures 4.2 and B.2). However, Figure D.1 shows that lexicon type 7 is the most informative one out of the three lexicon types of which the data is most confusable with that of type 14. It therefore makes sense that in a population that has converged on lexicon type 14 due to a selection pressure, learners continue to occasionally misinfer the data produced by their cultural parent as coming from a lexicon of type 7.

Lexicon type 7 ($ca = 0.45$) consists of all lexicons that are the exact *inverse* of the lexicons comprised in the most informative lexicon type 14. Lexicon type 5 ($ca = 0.4$) consists of all logically possible lexicons that are the same as those comprised in lexicon type 7, with the difference that one of the signals is associated with all three referents. Similarly, lexicon type 2 ($ca = 0.36$) consists of the same set of lexicons again, with the difference that two of the signals are associated with all three referents. Although these three lexicon types have almost identical values of confusability with the most informative lexicon type 14, only type 7 is represented in the populations under selection after convergence (see Figure 4.5 in Chapter 4). This is explained by the fact that agents who have received input from or have selected lexicon types 2 and 5 will have very little success under each of the three selection pressures explored in Chapter 4.

Appendix E

Confusability of lexicon types: pragmatic speakers

Figure E.1 shows how the confusability of lexicon types given the data they produce differs between literal and pragmatic speakers. This figure demonstrates that the overall confusability of data between lexicon types is much higher when it is produced by a pragmatic speaker than when it is produced by a literal speaker. As discussed in Chapter 5, this is a result of the fact that a pragmatic speaker will regularly use signals that are not strictly associated with their intended referent according to their lexicon, in order to avoid ambiguity elsewhere in the lexicon. A literal speaker on the other hand will only ever do so by 'mistake' (which has a relatively low probability of happening: $\epsilon = 0.05$). The higher level of confusability for pragmatic speakers explains the finding that pragmatic learners require more observations to correctly infer the lexicon of their cultural parent, as reported in Chapter 5 (Section 5.3.1). The confusability of lexicon types is similar between literal and pragmatic speakers in the sense that the highest levels of confusability are found between the lexicon types with lower levels of informativeness (types 1 through 7). The exact pattern of confusability however (i.e. which lexicon type is most confusable with which other lexicon types) differs between literal and pragmatic speakers.



(a) Literal speakers

(b) Pragmatic speakers

Figure E.1: Confusability matrices of lexicon types for literal and pragmatic speakers. SubFigure E.1a is the same as D.1, but rescaled according to the range of values shown in E.1b. Matrices show for each lexicon type how confusable the data it produces is with the data produced by other lexicon types. Note that the quantity shown in the matrices is not very interpretable given that it is the result of taking the inverse of the sum of the absolute difference between probabilities. We can however interpret the relative differences between the resulting values as relative differences in confusability (both within and between subfigures).

Appendix F

Modelling the co-development of word learning and perspective-taking

The following is a conference proceedings paper which was published in the Proceedings of the 38th Annual Meeting of the Cognitive Science Society in 2016, and is openly accessible. The modelling work presented in this paper forms a precursor to the modelling work presented in Chapter 3 of this thesis, and was conceived together with co-authors Simon Kirby, Chris Cummins and Kenny Smith. The paper was written by me, Marieke Woensdregt.

Modelling the co-development of word learning and perspective-taking

Marieke Woensdregt (m.s.woensdregt@sms.ed.ac.uk)[a]

Simon Kirby [a], **Chris Cummins** [b], **Kenny Smith** [a]

[a] Centre for Language Evolution, School of Philosophy, Psychology & Language Sciences, University of Edinburgh, 3 Charles Street, Edinburgh, EH8 9AD, United Kingdom

[b] Department of Linguistics and English Language, School of Philosophy, Psychology & Language Sciences, University of Edinburgh, 3 Charles Street, Edinburgh, EH8 9AD, United Kingdom

Abstract

Word learning involves mapping observable words to unobservable speaker intentions. The ability to infer referential intentions in turn has been shown to depend in part on access to language. Thus, word learning and intention-reading co-develop. To explore this interaction, we present an agent-based model in which an individual simultaneously learns a lexicon and learns about the speaker's perspective, given a shared context and the speaker's utterances, by performing Bayesian inference. Simulations with this model show that (i) lexicon-learning and perspective-learning are strongly interdependent: learning one is impossible without some knowledge of the other, (ii) lexicon- and perspective-learning can bootstrap each other, resulting in successful inference of both even when the learner starts with no knowledge of the lexicon and unhelpful assumptions about the minds of others, and (iii) receiving initial input from a 'helpful' speaker (who adopts the learner's perspective on the world) paves the way for later learning from speakers with perspectives which diverge from the learner's. This approach represents a first attempt to model the hypothesis that language and mindreading co-develop, and a first exploration of the implications for theories of word learning and mindreading development.

Keywords: word learning; perspective-taking; computational model; Bayesian inference;

Introduction

Word learning is a special case of associative learning, as one has to learn a mapping between something observable — a speaker's utterance — and something unobservable — the speaker's meaning. Word learning therefore requires inferring the speaker's referential intention (Waxman & Gelman, 2009), which in turn requires theory of mind (ToM). Learning about words and learning about minds are thus necessarily connected: language learners need to figure out not just the stable mappings between words and concepts (the lexicon) but also a way of inferring speaker intention, which is variable over time and depends on context and speaker-specific features.

In this paper we present evidence that language and ToM development go hand in hand, and we explore the implications of such a co-development by means of an agent-based model. As a test case we look specifically at the interaction between word learning and perspective-taking. Although perspective-taking cannot be equated with ToM, it is an instantiation of the latter and forms a good starting point for formalising the relation between language learning and ToM development.

Learning about words and minds

There is persuasive evidence consistent with the idea that learning about words and learning about minds are interrelated. In a study comparing children with autism (AD) to typically-developing (TD) children, Parish-Morris et al. (2007) showed that although 5-year-old AD children have some ability to use social cues (pointing and eye gaze) to direct their attention in word learning, they perform at chance when learning new words required inferring the speaker's intention, unlike language- and mental-age-matched TD children.

The reverse phenomenon has also been observed, namely that the development of ToM depends in part on having access to language. Deaf children of hearing parents, who lack consistent linguistic input, were shown to have delayed ToM development relative to deaf children of deaf parents, who receive sign language input from birth (Schick et al., 2007). Similarly, a study with TD children showed that simply training children on the use of mental state verbs with sentential complements accelerated their false belief understanding (Lohmann & Tomasello, 2003).

Thirdly, in a study comparing different age-groups of signers of the recently emerged Nicaraguan Sign Language, Pyers and Senghas (2009) showed that the bootstrap effect of language on ToM development continues on into adulthood. Pyers & Senghas found that the first cohort of signers (mean age 27), whose language had very limited mental state vocabulary, were worse at understanding false belief than the second cohort (mean age 17) who had more words for mental states. Moreover, a follow-up study two years later revealed that the first-cohort signers had improved in their false belief understanding and that this either followed or co-occurred with, but never preceded, an expansion of mental state vocabulary.

Finally, recent evidence suggests that mindreading and language skills co-develop. Brooks and Meltzoff (2015) showed that gaze-following in 10.5-month-old infants predicted their production of mental state terms at 2.5-years-old, and that these mental state terms in turn predicted the extent of their false belief understanding at 4.5-years-old, even though gaze-following did not directly predict false belief understanding. Thus, this shows evidence of an indirect relation between early sensitivity to social cues and later mindreading ability, mediated by language.

Models of word learning and perspective-taking

Words are used in complex environments, and each word could label any part of that complex environment. Worse, words can label objects and events which are not currently perceivable to the hearer and/or the speaker (e.g. events which are spatially or temporally distant from the time of speaking). Learners therefore face *referential uncertainty*: every time a word is used, there may be many meanings which a learner could infer as the word's intended meaning.

Computational models of word learning have explored several potential solutions to the problem of referential uncertainty, which could be roughly divided up into three kinds: (i) solutions using learning biases, (ii) social cues solutions, and (iii) intention-reading solutions.

Brute force statistical learning of word-referent associations is impossible if referential uncertainty is unbounded: if all logically possible meanings are equally-plausible candidates for the meaning of any word on any use, then no learner can learn the meaning of any word (an observation commonly attributed to Quine, 1960, in his work on radical translation). Experimental and observational studies have demonstrated that word learners use a number of heuristics to reduce referential uncertainty: learners assume that words refer to whole objects (Macnamara, 1972); they use argument structure and syntactic context to constrain the meaning of new words (Gillette et al., 1999); and they use knowledge of the meaning of other words to constrain hypotheses about the meaning of a new word, for example by assuming that words have mutually exclusive meanings (Markman & Wachtel, 1988). Models of cross-situational statistical learning suggest that brute-force cross-situational learning of large lexicons is possible under surprisingly high levels of referential uncertainty (Blythe, Smith, & Smith, 2010) or even under infinite referential uncertainty if word learners can use their heuristics to rank candidate meanings in terms of their plausibility (Blythe, Smith, & Smith, submitted).

In addition to exploiting linguistic context or their knowledge of likely word meanings, learners can use social cues, which are potentially highly informative in guiding word learning (see Paulus and Fikkert (2014) for eye-gaze and pointing and Yu and Smith (2012) for joint attention). Yu and Ballard (2007) formalised these mechanisms in a model of word learning that integrates the use of statistical regularities and social cues. They provided an associative model with information about which words and objects in a discourse stream were highlighted by social cues (prosody and joint attention), and simply increased the association weight of those items. They then tested the model on how well it could learn a lexicon from transcriptions of two videos of mother-child interactions from the CHILDES corpus. This 'hybrid' model was compared to a 'bare' statistical learning model, and statistical learners who exploited prosody or joint attention, but not both. Best performance was obtained with the model that integrated both types of social cue.

However, there is more to social interaction than just cues that direct attention. The ability to recognise that speech can convey unobservable communicative intentions comes online before children start talking (Vouloumanos, Onishi, & Pogue, 2012) and is used to guide their word learning (Parish-Morris et al., 2007). To formalise the role that inferring speaker intentions plays in word learning, Frank, Goodman, and Tenenbaum (2009) designed a Bayesian model that simultaneously infers word-object mappings and speaker intentions, and tested this model on the same CHILDES videos used by Yu and Ballard (2007). Rather than re-weighting items based on social cues, Frank et al. assume that learners posit an extra unobserved variable mediating between the objects in the physical context and the words that the speaker produces: the speaker's referential intention. The learner then evaluates all possible lexicon hypotheses based on the prior probability of that lexicon and the likelihood of a word given that lexicon and the speaker's referential intention, where the intention hypotheses that are considered by the learner are simply all possible subsets of the objects present in the context, including an 'empty' intention.

This model has two advantages over other associative learning models. Firstly, it can represent the possibility of 'empty intentions', where the word does not refer to any physically present object. Secondly, it can distinguish between words that can be used referentially and words that are used exclusively 'non-referentially', where non-referential (e.g. function) words are simply left out of the lexicon. Frank et al. (2009) show that this model outperforms several alternative statistical learning models (including Yu and Ballard's), both when tested on the lexicon they learned and on the referential intentions they inferred (given their lexicon).

Although these various models constitute important first steps towards modelling the role of intention-reading in word learning, they treat the ability to utilise social cues or infer intentions as a given and fixed capacity, present from the start of word learning. In real-world learning, the ability to learn words and the ability to infer mental states (including referential intentions) improve as a child grows older. As described above, this improvement is partly accounted for by a co-development of language and intention-reading. Below, we will describe a model that takes these considerations into account: rather than modelling word learning as a combination of associative learning with social cues or uninformed intention representations, we provide a model which allows for the co-development of word learning and perspective-taking.

The current model: Integrating development of word learning and perspective-taking

Model description

We model referential intentions as a result of the interaction between a set of attributes of the world — the context — and an attribute of the speaker — the perspective. This perspective can be interpreted in a literal sense, where objects that are spatially or temporally closer to the agent are more salient (see figure 1). Importantly however, it can equally serve as a

model for the sum of an agent's knowledge and beliefs about the world that determine what topics of conversation will be most salient to them in a given situation. The latter is the sort of perspective that requires full-blown ToM to be inferred. All that matters here is that there is a function that maps from the attributes of the world to an agent's saliency distribution over potential topics, and that the agent has a hidden variable (their perspective) that is a parameter in this function.

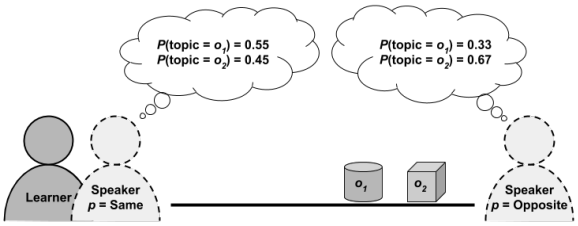


Figure 1: Diagram of how speaker perspective gives rise to referential intention. The speaker on the left is only slightly more likely to choose object 1 (o_1) over object 2 (o_2) as a referent, since both are approximately equidistant. The speaker on the right however is twice as likely to choose object 2 than object 1 since o_2 is twice as close as o_1. The learner has their own perspective on the world and learns with an egocentric bias; assuming that the speaker shares their perspective.

The variables that the learner can observe are the context and the speaker's utterance (see figure 2). The variables that are unobservable are the speaker's perspective, the speaker's referential intention, and the speaker's lexicon. The learner's task is to infer the speaker's perspective and the lexicon based on the same data: the speaker's word use in different contexts.

This model differs from that outlined in Frank et al. (2009) in that it posits an extra unobservable variable: the speaker's perspective, which together with the context determines the speaker's referential intention. Given a specific hypothesis about the speaker's perspective, the learner can compute a prediction of how likely it is that the speaker will refer to a given object in a given context (i.e. how salient the object is for the speaker). Subsequently, given a specific hypothesis about what the lexicon is, the learner can turn this prediction about likely referents into a prediction of likely utterances.

We assume, unlike the models of word learning described above, that all objects that are part of the world are possible referents in every learning context: thus, simple associative cross-situational learning alone will not be able to solve the problem of referential ambiguity. The learner can get around this problem by inferring the speaker's perspective: a hypothesis about this perspective is the only information available that can render the probability distribution over possible referents non-uniform, which in turn allows the learner to infer the most likely word-object mappings. Specifically, this is achieved by incrementing the posterior belief in lexicon hypotheses in proportion to how salient the object that is associated to the utterance in that lexicon is for the speaker, given the perspective hypothesis under consideration.

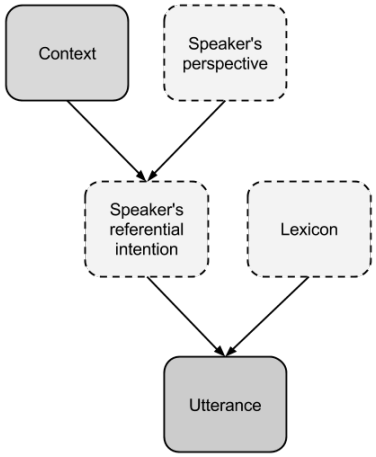


Figure 2: Diagram of the current model. Variables in dark grey and solid lines are observable to the learner, variables in light grey and dashed lines are unobservable. The learner's task is to infer the speaker's perspective and the lexicon based on observations of the speaker's word use in context.

Note that in this model no lexicon hypothesis can be evaluated without simultaneously positing a perspective hypothesis, and vice versa. Thus the complete hypothesis space for the learner consists of all possible combinations of lexicon hypothesis and perspective hypothesis (with the potential of representing different perspectives, and indeed different lexicons, for different speakers). Learning in this model is implemented as Bayesian inference according to the definitions described below.

Posterior The task of the learner in this model[1] is to find the lexicon hypothesis l and perspective hypothesis p that have the highest posterior probability given data D, as shown in equation 1.

$$P(l, p \mid D) \propto P(D \mid l, p)P(l, p) \quad (1)$$

The perspective hypothesis p represents a single parameter in an intention function that maps from the context to the speaker's referential intention. This referential intention is based on the saliency of the objects in the context, which is defined as the inverse of the distance between the speaker's perspective and the object's location (see figure 1). These saliency values are then normalized over all objects in the context, rendering a probability distribution over all objects

[1] We describe the model in terms of a learner who assumes that a single lexicon and a single speaker perspective will account for all of their data: the same model can straightforwardly be extended to model a learner who allows that different speakers might have different lexicons and different perspectives; later we present results for a learner who entertains multi-perspective hypotheses.

that determines how likely the speaker is to choose them as intended referent. This distribution is then used to generate the speakers referential intention.

The learner does not need to infer the intention function itself, only the perspective parameter. This model thus simulates the situation where the learner is 'born' with the ability to represent mental states, but has to learn how to make predictions about the *content* of another agent's mind on the basis of the context. More specifically, the learner is born with a model of how a context will give rise to a speaker's referential intention, given the speaker's perspective, but has to infer from data exactly what the speaker's perspective is.

Likelihood The likelihood of a set of data D is:

$$P(D \mid l,p) = \prod_{d \in D} P(w_d \mid l,p,c_d) \tag{2}$$

where each data point d consists of a context c and a word w that was uttered by the speaker in that particular context. The likelihood of a single word w_d is defined in equation 3.

$$P(w_d \mid l,p,c_d) = \sum_{o \in c_d} P(i_o \mid p,c_d) P(w_d \mid i_o,l) \tag{3}$$

where o stands for object and i_o for the probability that object o will be the intended referent given the perspective hypothesis p.

Thus, the probability of a particular word being uttered in a particular context is equal to the product of the probability of that word being uttered for a given object (according to lexicon hypothesis l) and the probability of that object being the intended referent (according to perspective hypothesis p), summed over all objects.

In the simulations described below all lexicon hypotheses that are considered consist simply of discrete binary mappings between words and objects — in other words, if there are two objects and two possible words, there are nine possible lexicons (object 1 maps to word a or word b or either, and object 2 independently maps to word a or word b or either). Thus, the probability of a given word being uttered for a given intended referent is given by equation 4

$$P(w_d \mid i_o,l) = \begin{cases} \dfrac{1}{|w_o|} \text{ if } w_d \text{ maps to } o \text{ in } l \\ 0 \text{ otherwise} \end{cases} \tag{4}$$

where $|w_o|$ is the number of words that map to object o in lexicon l.

Prior For all simulations described below, we assume that learners have a neutral prior over lexicons and an egocentric prior over perspectives. That is, the learner starts out assuming that all lexicons are equally probable, and that other agents share their own perspective. Over all combinations of lexicon and perspective prior, the 'composite prior' is simply the product of the two, as shown in equation 5.

$$P(l,p) = P(l)P(p) \tag{5}$$

Simulation results

All simulation results described in this section show what happens in the very simple case where the learner gets input from one or two speakers in a world where there exist only two possible referents (objects) and two words. The set of lexicon hypotheses consists of all functionally distinct ways of mapping two words onto two objects (nine lexicons in total, as described above). The set of perspective hypotheses consists of the two most extreme possibilities: either the speaker's perspective is the same as the learner's own perspective, or it is exactly the opposite. The learner's hypothesis space consists of all possible combinations of lexicon hypothesis and perspective hypothesis.

In a first set of four simulations we explore the influence that perspective-learning and lexicon-learning have on each other. We compare three different cases: (i) the target lexicon is unambiguous (i.e. each object is associated with a distinct word) but the learner is unable to learn that speakers might have a perspective that is different from their own (which we achieve by setting the prior probability of the 'other' perspective to 0); (ii) the learner is initially egocentric yet can learn that speakers can have a perspective that differs from their own (which we achieve by setting the prior probability of the 'other' perspective to 0.1, and the 'own' perspective to 0.9), but the target lexicon is partly ambiguous (e.g. object 1 maps to both word a and word b, while object 2 maps only to word b); (iii) same as in (ii) but with a fully ambiguous lexicon (both objects map to both words); and (iv) the learner can learn that the speakers can have a different perspective from the learner, as in (ii) and (iii), and the target lexicon is unambiguous, as in (i).

Situation (iv) thus simulates a typically-developing child in a normal language environment (under the assumption that words are effectively unambiguous in their linguistic context: Piantadosi, Tily, & Gibson, 2012) — we refer to this as the Typical condition. Situation (i) simulates a word learner with a strongly impaired (or absent) ToM — we refer to this as the No ToM condition. Situation (ii), which we refer to as the Partly Ambiguous Lexicon condition, simulates a typically-developing word learner in an environment where the target lexicon is such that a speaker's utterances are rather uninformative about their referential intentions. This scenario could be compared to the case of deaf children who grow up with hearing parents (i.e. without sign language), since although such parents do exhibit communicative behaviour that could reveal something about their communicative intentions, this is less explicit and more ambiguous than linguistic data (Schick et al., 2007). Finally, situation (iii), which we refer to as the Uninformative Lexicon condition, is an extreme form of this case, where there is a complete absence of behaviour that is informative about the speaker's intentions. This is a case analogous to one in which a reliable language has yet to emerge in a population.

Figure 3 shows the learning results for the four different situations described above. Several interesting learning dynam-

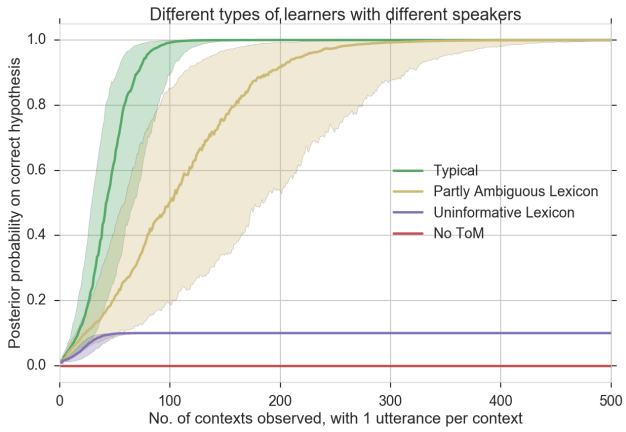


Figure 3: Learning curves for different learners in different learning situations. Learning is measured as the amount of posterior probability assigned to the correct hypothesis, where 1.0 is ceiling. Lines show median over 1000 runs, shaded area shows first and third quartile.

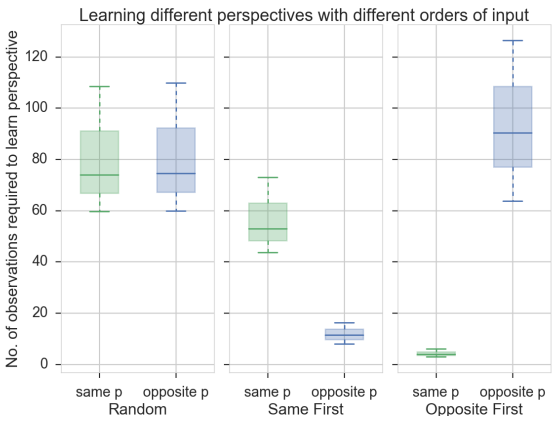


Figure 4: Amount of observations required for learning different speaker perspectives under different input conditions: Random, Same First and Opposite First. Successful learning is defined as > 0.99 posterior probability on correct hypothesis, and the lexicon is learned fully in all conditions before the learner enters the second input phase. Boxes show median, first and third quartile over 100 runs.

ics are apparent. Firstly, inferring the correct lexicon is impossible when the learner cannot infer the correct perspective of the speaker that they get input from (No ToM condition). Secondly, inferring the speaker's perspective becomes more difficult when there is a less direct mapping between their referential intention and their behaviour (Partly Ambiguous Lexicon condition). However, learning in this case is still eventually successful: the ability to infer perspective gives a way into learning the lexicon, thus making it easier to deal with lexical ambiguity. Thirdly, inferring the speaker's perspective becomes impossible when the speaker's behaviour gives no information at all about their intention (Uninformative Lexicon condition). Finally, learning happens most quickly and successfully when the learner is both able to represent different perspectives and the speakers' lexicon is unambiguous (Typical condition).[2]

In a second set of three simulations we present the effect of order of input on lexicon and perspective learning. These simulations are similar to the ones described above, except that the learner receives input from two different speakers who have two different perspectives: one speaker shares the learner's perspective, the other has the opposite perspective. We present the learning results in three different situations: (i) the speaker is randomly picked on each trial, but both speakers get to speak for an equal number of contexts (Random condition); (ii) the learner receives the first half of their input from the speaker that shares their perspective, and the second half from the 'opposite perspective' speaker (Same First condition); and (iii) the learner receives the first half of input from the opposite perspective speaker and the second half from the same perspective speaker (Opposite First condition).

[2]These results are qualitatively similar for learning about larger lexicons of e.g. 3x3 and 4x4 objects and words.

As figure 4 shows, the difference in the amount of observations that is required to learn the opposite perspective is bigger between the two conditions (Same First vs. Opposite First) than the difference in the amount of observations required to learn the same perspective in the two conditions. This means that receiving input from a 'helpful' speaker (a speaker who shares the learner's perspective) first paves the way for later learning about perspectives that are different from the learner's own.[2]

The mediating factor that gives rise to this effect is the lexicon, since the only thing that is different about the learner after having learned the same perspective first is their knowledge of the lexicon. (Which, in all simulations shown in figure 4, is fully learned before the learner enters the second input phase.) This effect relies on the lexicon being shared among members of the population. Language as a convention is what allows the learner to bootstrap knowledge of other's perspectives based on starting with a familiar speaker first.

Discussion

We presented an agent-based model that simulates the co-development of word-learning and perspective-taking through Bayesian inference. This model is different from previous models of word learning in that all objects that are part of the world are considered as potential referents at each learning episode, rendering brute-force cross-situational learning impossible. However, the learner can overcome this referential uncertainty by learning about the speaker's perspective. Both the lexicon and the perspective are learned using the same data (the speaker's word use in context).

This model gives rise to several potentially interesting co-development dynamics. Firstly, lexicon-learning and

perspective-learning are strongly interdependent: learning the one cannot happen without some knowledge of the other. Secondly, lexicon- and perspective-learning can bootstrap each other, resulting in successful inference of both variables even when the learner starts out with an inappropriate egocentric bias and no knowledge of the lexicon whatsoever. Finally, the results show that receiving input from a helpful speaker first paves the way for later learning from speakers whose perspective differs from the learner's — the helpful speaker provides data which facilitates learning of the lexicon, which then facilitates learning of the perspective of other less well-aligned speakers (on the assumption that the lexicon is shared among speakers).

To our knowledge, this is the first computational model that does not simply incorporate pragmatic inference as a tool to infer word meaning (Frank et al., 2009), but rather incorporates pragmatic inference as a developing skill that interacts bi-directionally with word learning. Thus, this model is a first step towards formalising the hypothesis that language and mindreading co-develop.

The simulation results of this model described here replicate several empirical findings. Firstly, it mirrors the finding that word-learning depends partly on the inference of mental states (Parish-Morris et al., 2007). Secondly, it mirrors the finding that the development of mindreading depends partly on vocabulary development (Lohmann & Tomasello, 2003; Pyers & Senghas, 2009; Schick et al., 2007). Finally, it generates the developmental prediction that learning from a helpful speaker who shares the child's perspective early on in life will aid vocabulary development, and that this in turn will help the child to learn about alternative perspectives later on.

Several aspects of this model are however very simplistic. Firstly, the learner in this model is 'born' with a ToM. Rather than having to infer the full function that maps from a context to a speaker's referential intention, the learner only has to infer the speaker's perspective. In real life children have to develop not only the ability to infer the content of mental states, but also the underlying ability to represent *that* the content of others' minds is different from that of their own. Future work with this model could incorporate a more realistic model of ToM development that could mimic more closely the stages of ToM development we see in real children.

Secondly, the relation between observations of words and learning about perspectives is very direct. Each word-object mapping that is learned helps with inferring perspective because it allows the learner to evaluate their prediction of referential intent based on their perspective hypothesis. It is not yet clear what the role of language learning is in driving the development of ToM in the real world — this might have to do with access to discourse, explanations or representations of mental states (see e.g. Lohmann & Tomasello, 2003; Pyers & Senghas, 2009; Schick et al., 2007).

Despite these simplifications, this model forms a first exploration into the co-development dynamics of language and ToM.

References

Blythe, R. A., Smith, A. D. M., & Smith, K. (submitted). Word learning under infinite uncertainty.

Blythe, R. A., Smith, K., & Smith, A. D. M. (2010). Learning times for large lexicons through cross-situational learning. *Cognitive Science*, *34*, 620–42.

Brooks, R., & Meltzoff, A. N. (2015). Connecting the dots from infancy to childhood: A longitudinal study connecting gaze following, language, and explicit theory of mind. *Journal of Experimental Child Psychology*, *130*, 67–78.

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), 578–85.

Gillette, J., Gleitman, H., Gleitman, L., & Lederer, A. (1999). Human simulations of vocabulary learning. *Cognition*, *73*, 135–76.

Lohmann, H., & Tomasello, M. (2003). The Role of Language in the Development of False Belief Understanding: A Training Study. *Child Development*, *74*(4), 1130–44.

Macnamara, J. (1972). The cognitive basis of language learning in infants. *Psychological Review*, *79*, 1–13.

Markman, E. M., & Wachtel, G. F. (1988). Children's use of mutual exclusivity to constrain the meaning of words. *Cognitive Psychology*, *20*, 121–57.

Parish-Morris, J., Hennon, E. a., Hirsh-Pasek, K., Golinkoff, R. M., & Tager-Flusberg, H. (2007). Children with autism illuminate the role of social intention in word learning. *Child Development*, *78*(4), 1265–87.

Paulus, M., & Fikkert, P. (2014). Conflicting Social Cues: Fourteen- and 24-Month-Old Infants' Reliance on Gaze and Pointing Cues in Word Learning. *Journal of Cognition and Development*, *15*(1), 43–59.

Piantadosi, S. T., Tily, H., & Gibson, E. (2012). The communicative function of ambiguity in language. *Cognition*, *122*, 280-91.

Pyers, J. E., & Senghas, A. (2009). Language Promotes False-Belief Understanding: Evidence From Learners of a New Sign Language. *Psychological Science*, *20*, 805–12.

Quine, W. V. O. (1960). *Word and Object*.

Schick, B., de Villiers, P., de Villiers, J., & Hoffmeister, R. (2007). Language and theory of mind: a study of deaf children. *Child Development*, *78*(2), 376–96.

Vouloumanos, A., Onishi, K. H., & Pogue, A. (2012). Twelve-month-old infants recognize that speech can communicate unobservable intentions. *PNAS*, *109*, 12933–7.

Waxman, S. R., & Gelman, S. a. (2009). Early word-learning entails reference, not merely associations. *Trends in Cognitive Sciences*, *13*(6), 258–63.

Yu, C., & Ballard, D. H. (2007). A unified model of early word learning: Integrating statistical and social cues. *Neurocomputing*, *70*, 2149–65.

Yu, C., & Smith, L. B. (2012). Embodied attention and word learning by toddlers. *Cognition*, *125*(2), 244–62.

References

Acerbi, A. and Mesoudi, A. (2015). If we are all cultural Darwinians what's the fuss about? Clarifying recent disagreements in the field of cultural evolution. *Biology and Philosophy*, 30:481–503.

Akhtar, N. and Montague, L. (1999). Early lexical acquisition: The role of cross-situational learning. *First Language*, 19(57):347–358.

Akhtar, N. and Tomasello, M. (1996). Two-year-olds learn words for absent objects and actions. *British Journal of Developmental Psychology*, 14(1):79–93.

Ale, S. B., Brown, J. S., and Sullivan, A. T. (2013). Evolution of Cooperation: Combining Kin Selection and Reciprocal Altruism into Matrix Games with Social Dilemmas. *PLOS ONE*, 8(5):1–9.

Anderson, D. K., Lord, C., Risi, S., DiLavore, P. S., Shulman, C., Thurm, A., Welch, K., and Pickles, A. (2007). Patterns of growth in verbal abilities among children with autism spectrum disorder. *Journal of Consulting and Clinical Psychology*, 75(4):594–604.

Apperly, I. (2011). *Mindreaders: The Cognitive Basis of "Theory of Mind"*. Psychology Press.

Apperly, I. A. and Butterfill, S. A. (2009). Do humans have two systems to track beliefs and belief-like states? *Psychological Review*, 116(4):953–70.

Association, A. P. (2013). *Diagnostic and Statistical Manual of Mental Disorders (5th Ed.)*. American Psychiatric Publishing, Arlington, VA, 5 edition.

Astington, J. W. and Baird, J. A., editors (2005). *Why Language Matters for Theory of Mind*. Why Language Matters for Theory of Mind. Oxford University Press, New York, NY, US.

Baixauli, I., Colomer, C., Roselló, B., and Miranda, A. (2016). Narratives of children with high-functioning autism spectrum disorder: A meta-analysis. *Research in Developmental Disabilities*, 59:234–254.

Baldwin, D. A. (1991). Infants' contribution to the achievement of joint reference. *Child Dev*, 62(5):875–890.

Baldwin, D. A. (1993a). Early referential understanding: Infants' ability to recognize referential acts for what they are. *Developmental Psychology*, 29(5):832–843.

Baldwin, D. A. (1993b). Infants' ability to consult the speaker for clues to word reference. *J Child Lang*, 20(2):395–418.

Baldwin, D. A., Markman, E. M., Bill, B., Desjardins, R. N., Irwin, J. M., and Tidball, G. (1996). Infants' reliance on a social criterion for establishing word-object relations. *Child development*, 67(6):3135–3153.

Baldwin, D. A. and Moses, L. J. (2001). Links between Social Understanding and Early Word Learning: Challenges to Current Accounts. *Social Development*, 10(3):309–329.

Bar-On, D. (2013). Origins of Meaning: Must We 'Go Gricean'? *Mind & Language*, 28(3):342–375.

Bar-On, D. (2016). Sociality, Expression, and This Thing called Language. *Inquiry*, 59(1):56–79.

Baron-Cohen, S. (1988). Social and pragmatic deficits in autism: Cognitive or affective? *J Autism Dev Disord*, 18(3):379–402.

Baron-Cohen, S. (1995). *Mindblindness: An Essay on Autism and Theory of Mind*, volume 74. MIT Press. Publication Title: Learning development and conceptual change.

Baron-Cohen, S., Baldwin, D. A., and Crowson, M. (1997). Do Children with Autism Use the Speaker's Direction of Gaze Strategy to Crack the Code of Language? *Child Development*, 68(1):48–57.

Baron-Cohen, S., Jolliffe, T., Mortimore, C., and Robertson, M. (1997). Another Advanced Test of Theory of Mind: Evidence from Very High Functioning Adults with Autism or Asperger Syndrome. *Journal of Child Psychology and Psychiatry*, 38(7):813–822.

Baron-Cohen, S., Leslie, A. M., and Frith, U. (1985). Does the autistic child have a "theory of mind"? *Cognition*, 21:37–46.

Barrett, H. C., Broesch, T., Scott, R. M., He, Z., Baillargeon, R., Wu, D., Bolz, M., Henrich, J., Setoh, P., Wang, J., and Laurence, S. (2013). Early false-belief understanding in traditional non-Western societies. *Proceedings of the Royal Society B*, 280(1755):20122654.

Barton, N. H., Briggs, D. E., Eisen, J. A., Goldstein, D. B., and Patel, N. H. (2007). *Evolution*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

Bello, P. (2012). Cognitive foundations for a computational theory of mindreading. *Advances in Cognitive Systems*, 1:59–72.

Beran, M. J., Smith, J. D., and Perdue, B. M. (2013). Language-trained chimpanzees (Pan troglodytes) name what they have seen, but look first at what they have not seen. *Psychological Science*, 24(5):660–666.

Birch, S. A. J. and Bloom, P. (2004). Understanding children's and adults' limitations in mental state reasoning. *Trends in Cognitive Sciences*, 8(6):255–260.

Blythe, R. A., Smith, A. D., and Smith, K. (2016). Word learning under infinite uncertainty. *Cognition*, 151:18–27.

Blythe, R. A., Smith, K., and Smith, A. D. M. (2010). Learning times for large lexicons through cross-situational learning. *Cognitive science*, 34(4):620–42.

Boyd, R. and Richerson, P. J. (1985). *Culture and the Evolutionary Process*. University of Chicago Press.

Breheny, R. (2006). Communication and Folk Psychology. *Mind & Language*, 21(1):74–107.

Brighton, H. (2002). Compositional syntax from cultural transmission. *Artificial Life*, 8(1):25–54.

Brighton, H., Smith, K., and Kirby, S. (2005). *Language as an Evolutionary System*, volume 2. Publication Title: Physics of Life Reviews.

Brochhagen, T., Franke, M., and van Rooij, R. (2018). Coevolution of Lexical Meaning and Pragmatic Use. *Cognitive Science*, 0(0).

Brooks, R. and Meltzoff, A. N. (2015). Connecting the dots from infancy to childhood: A longitudinal study connecting gaze following, language, and explicit theory of mind. *Journal of Experimental Child Psychology*, 130:67–78.

Brüne, M. and Brüne-Cohrs, U. (2006). Theory of mind-evolution, ontogeny, brain mechanisms and psychopathology. *Neuroscience and Biobehavioral Reviews*, 30(4):437–455.

Burkart, J. M., Hrdy, S. B., and Van Schaik, C. P. (2009). Cooperative breeding and human cognitive evolution. *Evolutionary Anthropology*, 18(5):175–186.

Burkett, D. and Griffiths, T. L. (2010). Iterated learning of multiple languages from multiple teachers. In *The Evolution of Language: Proceedings of the 8th International Conference (EVOLANG8), Utrecht, Netherlands, 14-17 April 2010*, pages 58–65.

Butterfill, S. A. and Apperly, I. A. (2013). How to construct a minimal theory of mind. *Mind and Language*, 28(5):606–637.

Byrne, D. (1996). Machiavellian Intelligence II. *Evolutionary Anthropology*, 5(5):172–180.

Call, J. (2010). Do apes know that they could be wrong? *Animal Cognition*, 13:689–700.

Call, J. and Carpenter, M. (2000). Do apes and children know what they have seen? *Animal Cognition*, 4:207–220.

Call, J. and Tomasello, M. (2008). Does the chimpanzee have a theory of mind? 30 years later. *Trends in Cognitive Sciences*, 12(5):187–192.

Carpenter, M., Nagell, K., and Tomasello, M. (1998). Social cognition, joint attention, and communicative competence from 9 to 15 months of age. *Monogr Soc Res Child Dev*, 63(4):i–vi, 1–143.

Carston, R. (2002). *Thoughts and Utterances*. Blackwell Publishing Ltd.

Charman, T., Baron-Cohen, S., Swettenham, J., Baird, G., Cox, A., and Drew, A. (2000). Testing joint attention, imitation, and play as infancy precursors to language and theory of mind. *Cognitive Development*, 15(4):481–498.

Claidière, N., Amedon, G. K.-k., André, J.-B., Kirby, S., Smith, K., Sperber, D., and Fagot, J. (2018). Convergent transformation and selection in cultural evolution. *Evolution and Human Behavior*, 39(2):191–202.

Claidière, N., Scott-Phillips, T. C., and Sperber, D. (2014a). How Darwinian is cultural evolution? *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 369(1642):20130368.

Claidière, N., Smith, K., Kirby, S., and Fagot, J. (2014b). Cultural evolution of systematically structured behaviour in a non-human primate. *Proceedings. Biological sciences / The Royal Society*, 281(1797):20141541.

Claidière, N. and Sperber, D. (2007). The role of attraction in cultural evolution. *Journal of Cognition and Culture*, 7(1):89–111.

Crockford, C., Wittig, R. M., Mundry, R., and Zuberbühler, K. (2012). Wild Chimpanzees Inform Ignorant Group Members of Danger. *Current Biology*, 22:142–146.

Csibra, G. (2008). Goal attribution to inanimate agents by 6.5-month-old infants. *Cognition*, 107(2):705–17.

Csibra, G. (2010). Recognizing Communicative Intentions in Infancy. *Mind & Language*, 25(2):141–168.

Csibra, G. and Gergely, G. (2006). Social learning and social cognition: The case for pedagogy. In Munakata, Y. and Johnson, M. H., editors, *Processes of Change in Brain and Cognitive Devolpment. Attention and Performance, XXI*, pages 249–274. Oxford University Press.

Csibra, G. and Gergely, G. (2011). Natural pedagogy as evolutionary adaptation. *Philosophical Transactions of the Royal Society B*, 366(1567):1149–1157.

De Giacomo, A. and Fombonne, E. (1998). Parental recognition of developmental abnormalities in autism. *Eur Child Adolesc Psychiatry*, 7(3):131–136.

de Marchena, A., Eigsti, I.-M., Worek, A., Ono, K. E., and Snedeker, J. (2011). Mutual exclusivity in autism spectrum disorders: Testing the pragmatic hypothesis. *Cognition*, 119(1):96–113.

de Villiers, J. (2007). The interface of language and Theory of Mind. *Lingua*, 117:1858–1878.

de Villiers, J. G. and Pyers, J. E. (2002). Complements to cognition: A longitudinal study of the relationship between complex syntax and false-belief-understanding. *Cognitive Development*, 17:1037–1060.

de Villiers, P. A. and de Villiers, J. G. (2012). Deception dissociates from false belief reasoning in deaf children: Implications for the implicit versus explicit theory of mind distinction. *British Journal of Developmental Psychology*, 30(1):188–209.

Dennett, D. C. (1983). Intentional systems in cognitive ethology: The "Panglossian paradigm" defended. *The Behavioral and Brain Sciences*, 6:343–390.

Dunstone, J. and Caldwell, C. A. (2018). Cumulative culture and explicit metacognition: A review of theories, evidence and key predictions. *Palgrave Communications*, 4(1):145.

Eigsti, I.-M., de Marchena, A. B., Schuh, J. M., and Kelley, E. (2011). Language acquisition in autism spectrum disorders: A developmental review. *Research in Autism Spectrum Disorders*, 5(2):681–691.

Flavell, J. H., Green, F. L., and Flavell, E. R. (1990). Developmental changes in young children's knowledge about the mind. *Cognitive Development*, 5(1):1–27.

Frank, M. and Goodman, N. (2014). Inferring word meanings by assuming that speakers are informative. *Cognitive Psychology*, 75:80–96.

Frank, M. C., Emilsson, A. G., Peloquin, B., Goodman, N. D., and Potts, C. (2017). Rational speech act models of pragmatic reasoning in reference games. Publication Title: PsyArxiv.

Frank, M. C. and Goodman, N. D. (2012). Predicting Pragmatic Reasoning in Language Games. *Science*, 336:998.

Frank, M. C., Goodman, N. D., Lai, P., and Tenenbaum, J. B. (2009a). Informative communication in word production and word learning. *Proceedings of the 31St Annual Conference of the Cognitive Science Society*, pages 1228–1233.

Frank, M. C., Goodman, N. D., and Tenenbaum, J. B. (2009b). Using speakers' referential intentions to model early cross-situational word learning. *Psychological science*, 20(5):578–85.

Frank, M. C., Tamnes, C. K., Reschke, P. J., Rocha-Hidalgo, J., and Lieberman, A. (2018). ManyBabies 2: Infant Theory of Mind. Retrieved from osf.io/jmuvd.

Franke, M. (2017). Game Theory in Pragmatics: Evolution, Rationality & Reasoning. In *Oxford Research Encyclopedia of Linguistics*, pages 1–23.

Franke, M. and Degen, J. (2016). Reasoning in Reference Games: Individual- vs. Population-Level Probabilistic Modeling. *PLOS ONE*, 11(5):e0154854.

Franke, M. and Jäger, G. (2014). Pragmatic Back-and-Forth Reasoning. In Reda, S. P., editor, *Pragmatics, Semantics and the Case of Scalar Implicatures*, pages 170–200. Palgrave MacMillan, New York.

Franke, M. and Jäger, G. (2016). Probabilistic pragmatics, or why Bayes' rule is probably important for pragmatics. *Zeitschrift fur Sprachwissenschaft*, 35(1):3–44.

Frith, U. and Happé, F. (1994). Autism: Beyond "theory of mind". *Cognition*, 50(1):115–132.

Gagne, D. L. and Coppola, M. (2017). Visible Social Interactions Do Not Support the Development of False Belief Understanding in the Absence of Linguistic Input: Evidence from Deaf Adult Homesigners. *Frontiers in Psychology*, 8(June):1–21.

Gärdenfors, P. (2003). *How Homo Became Sapiens: On the Evolution of Thinking*. Oxford University Press.

Gleitman, L. (1990). The Structural Sources of Verb Meanings. *Language Acquisition*, 1(1):3–55.

Gliga, T., Elsabbagh, M., Hudry, K., Charman, T., and Johnson, M. H. (2012). Gaze Following, Gaze Reading, and Word Learning in Children at Risk for Autism. *Child Development*, 83(3):926–938.

Gómez, J.-C. (1994). Mutual awareness in primate communication: A Gricean approach. In Parker, S., Boccia, M., and Mitchell, R., editors, *Self-Recognition and Awareness in Apes, Monkeys and Children*, pages 61–80. Cambridge, UK: Cambridge University Press.

Gómez, J. C. (2007). Pointing behaviors in apes and human infants: A balanced interpretation. *Child Development*, 78(3):729–734.

Gong, T. and Shuai, L. (2012). Modelling the coevolution of joint attention and language. *Proceedings. Biological sciences / The Royal Society*, 279(1747):4643–51.

Goodman, N. D. and Frank, M. C. (2016). Pragmatic Language Interpretation as Probabilistic Inference. *Trends in Cognitive Sciences*, 20(11):818–829.

Goodman, N. D. and Stuhlmüller, A. (2013). Knowledge and Implicature: Modeling Language Understanding as Social Cognition. *Topics in Cognitive Science*, 5(1):173–184.

Gould, S. J. and Vrba, E. S. (1982). Exaptation - A Missing Term in the Science of Form. *Paleobiology*, 8(1):4–15.

Grice, H. P. (1957). Meaning. *The Philosophical Review*, 66(3):377–388.

Grice, H. P. (1975). Logic and Conversation. In Grice, H. P., editor, *Studies in the Way of Words*, pages 305–315. Harvard University Press.

Griffiths, T. L. and Kalish, M. L. (2007). Language evolution by iterated learning with bayesian agents. *Cognitive science*, 31:441–480.

Griffiths, T. L., Kalish, M. L., and Lewandowsky, S. (2008). Review. Theoretical and empirical evidence for the impact of inductive biases on cultural evolution. *Philosophical Transactions of the Royal Society B-Biological Sciences: Biological Sciences*, 363(1509):3503–3514.

Happé, F. (1999). Autism: Cognitive deficit or cognitive style? *Trends in Cognitive Sciences*, 3(6):216–222.

Happé, F. and Frith, U. (2006). The Weak Coherence Account: Detail-focused Cognitive Style in Autism Spectrum Disorders. *J Autism Dev Disord*, 36(1):5–25.

Happé, F., Ronald, A., and Plomin, R. (2006). Time to give up on a single explanation for autism. *Nature Neuroscience*, 9(10):1218–1220.

Harris, P. L. (1996). Desires, beliefs, and language. In Carruthers, P. and Smith, P. K., editors, *Theories of Theories of Mind*, pages 200–220. Cambridge University Press.

Helming, K. A., Strickland, B., and Jacob, P. (2014). Making sense of early false-belief understanding. *Trends in Cognitive Sciences*, 18(4):167–170.

Henrich, J. and Boyd, R. (2002). On Modeling Cognition and Culture representations. *Journal of Cognition and Culture*, 2(2):87–112.

Henrich, J., Boyd, R., and Richerson, P. J. (2008). Five misunderstandings about cultural evolution. *Human Nature*, 19(2):119–137.

Henrich, J. and McElreath, R. (2003). The Evolution of Cultural Evolution. *Evolutionary Anthropology*, 12(3):123–135.

Herrmann, E., Call, J., Hernández-Lloreda, M. V., Hare, B., and Tomasello, M. (2007). Humans Have Evolved Specialized Skills of Social Cognition: The Cultural Intelligence Hypothesis. *Science*, 1360(317):1360–1366.

Heyes, C. (2012a). Grist and mills: On the cultural origins of cultural learning. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 367(1599):2181–91.

Heyes, C. (2012b). New thinking: The evolution of human cognition. *Philosophical Transactions of the Royal Society B*, 367:2091–2096.

Heyes, C. (2012c). What's social about social learning? *Journal of Comparative Psychology*, 126(2):193–202.

Heyes, C. (2014a). False belief in infancy: A fresh look. *Developmental science*, 17(5):647–659.

Heyes, C. (2014b). Submentalizing: I Am Not Really Reading Your Mind. *Perspectives on Psychological Science*, 9(2):131–143.

Heyes, C. (2015). Animal mindreading: What's the problem? *Psychonomic Bulletin and Review*, 22(2):313–327.

Heyes, C. (2018). *Cognitive Gadgets*. Harvard University Press.

Heyes, C. M. and Frith, C. D. (2014). The cultural evolution of mind reading. *Science*, 344(6190).

Hill, E. L. (2004). Executive dysfunction in autism. *Trends in Cognitive Sciences*, 8(1):26–32.

Hobson, R. P., García-Pérez, R. M., and Lee, A. (2010a). Person-Centred (Deictic) Expressions and Autism. *Journal of Autism and Developmental Disorders*, 40(4):403–415.

Hobson, R. P., Lee, A., and a. Hobson, J. (2010b). Personal pronouns and communicative engagement in autism. *Journal of Autism and Developmental Disorders*, 40:653–664.

Hofmann, S. G., Doan, S. N., Sprung, M., Wilson, A., Ebesutani, C., Andrews, L. A., Curtiss, J., and Harris, P. L. (2016). Training children's theory-of-mind: A meta-analysis of controlled studies. *Cognition*, 150:200–212.

Howlin, P. (2003). Outcome in High-Functioning Adults with Autism with and Without Early Language Delays: Implications for the Differentiation Between Autism and Asperger Syndrome. page 11.

Hughes, C., Jaffee, S. R., Happé, F., Taylor, A., Caspi, A., and Moffitt, T. E. (2005). Origins of Individual Differences in Theory of Mind: From Nature to Nurture? *Child Development*, 76(2):356–370.

Kachergis, G., Yu, C., and Shiffrin, R. M. (2012). An associative model of adaptive inference for learning word–referent mappings. *Psychonomic Bulletin & Review*, 19(2):317–324.

Kaminski, J., Call, J., and Tomasello, M. (2008). Chimpanzees know what others know, but not what they believe. *Cognition*, 109:224–234.

Kampis, D., Somogyi, E., Itakura, S., and Király, I. (2013). Do infants bind mental states to agents? *Cognition*, 129:232–240.

Kao, J. T., Bergen, L., and Goodman, N. D. (2014a). Formalizing the Pragmatics of Metaphor Understanding. *Proceedings of the 36th Annual Conference of the Cognitive Science Society (CogSci 2014)*, 1:719–724.

Kao, J. T. and Goodman, N. D. (2015). Let's talk (ironically) about the weather: Modeling verbal irony. *Proceedings of the 36th Conference of the Cognitive Science Society*, pages 1051–1056.

Kao, J. T., Wu, J. Y., Bergen, L., and Goodman, N. D. (2014b). Nonliteral understanding of number words. *Proceedings of the National Academy of Sciences of the United States of America*, (7):1–6.

Kasari, C., Gulsrud, A., Freeman, S., Paparella, T., and Hellemann, G. (2012). Longitudinal Follow Up of Children with Autism Receiving Targeted Interventions on Joint Attention and Play RH = Targeted Interventions on Joint Attention and Play. *J Am Acad Child Adolesc Psychiatry*, 51(5):487–495.

Kazak, S., Collis, G. M., and Lewis, V. (1997). Can young people with autism refer to knowledge states? Evidence from their understanding of "know" and "guess". *J Child Psychol Psychiatry*, 38(8):1001–1009.

Kelley, E., Paul, J. J., Fein, D., and Naigles, L. R. (2006). Residual Language Deficits in Optimal Outcome Children with a History of Autism. *J Autism Dev Disord*, 36(6):807.

Kemp, C., Perfors, A., and Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental science*, 10(3):307–21.

Kirby, S. (2000). Syntax Without Natural Selection: How Compositionality Emerges from Vocabulary in a Population of Learners. In Knight, C., Studdert-Kennedy, M., and Hurford, J., editors, *The Evolutionary Emergence of Language*, pages 303–323. Cambridge University Press, Cambridge.

Kirby, S. (2001). Spontaneous evolution of linguistic structure-an iterated learning model of the emergence of regularity and irregularity. *IEEE Transactions on Evolutionary Computation*, 5(2):102–110.

Kirby, S. (2002). Learning, bottlenecks and the evolution of recursive syntax. In Briscoe, T., editor, *Linguistic Evolution through Language Acquisition: Formal and Computational Models*, pages 173–204. Cambridge University Press.

Kirby, S. (2017). Culture and biology in the origins of linguistic structure. *Psychonomic Bulletin and Review*, 24(1):118–137.

Kirby, S., Dowman, M., and Griffiths, T. L. (2007). Innateness and culture in the evolution of language. *Proceedings of the National Academy of Sciences of the United States of America*, 104(March):5241–5245.

Kirby, S., Griffiths, T., and Smith, K. (2014). Iterated learning and the evolution of language. *Current Opinion in Neurobiology*, 28:108–114.

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

Kovács, A. M., Téglás, E., and Endress, A. D. (2010). The social sense: Susceptibility to others' beliefs in human infants and adults. *Science*, 330(6012):1830–4.

Krachun, C., Carpenter, M., Call, J., and Tomasello, M. (2009). A competitive nonverbal false belief task for children and apes. *Developmental Science*, 12(4):521–35.

Krupenye, C., Kano, F., Hirata, S., Call, J., and Tomasello, M. (2016). Great apes anticipate that other individuals will act according to false beliefs. *Science*, 354(6308):110–114.

Kwisthout, J., Vogt, P., Haselager, P., and Dijkstra, T. (2008). Joint attention and language evolution. *Connection Science*, 20(2-3):155–171.

Lachlan, R. F. and Slater, P. J. B. (1999). The maintenance of vocal learning by gene–culture interaction: The cultural trap hypothesis. *Proceedings of the Royal Society of London B: Biological Sciences*, 266(1420):701–706.

Leavens, D. A., Russell, J. L., and Hopkins, W. D. (2010). Multimodal communication by captive chimpanzees (Pan troglodytes). *Animal Cognition*, 13(1):33–40.

Leslie, A. M. (2005). Developmental parallels in understanding minds and bodies. *Trends in Cognitive Sciences*, 9(10):459–62.

Leslie, A. M., Friedman, O., and German, T. P. (2004). Core mechanisms in "theory of mind". *Trends in Cognitive Sciences*, 8(12):528–33.

Levinson, S. C. (2000). *Presumptive Meanings: The Theory of Generalized Conversational Implicature Stephen*. MIT Press.

Lewis, D. (1969). *Convention: A Philosophical Study*. Harvard University Press.

Liebal, K., Call, J., Tomasello, M., Pika, S., Call, J., and Tomasello, M. (2004). To move or not to move: How apes adjust to the attentional state of others. *Interaction Studies*, 5(2):199–219.

Liebal, K., Waller, B. M., Burrows, A. M., and Slocombe, K. E. (2014). *Primate Communication: A Multimodal Approach*. Cambridge University Press.

Liittschwager, J. C. and Markman, E. M. (1994). Sixteen- and 24-month-olds' use of mutual exclusivity as a default assumption in second-label learning. *Developmental Psychology*, 30(6):955–968.

Lillard, A. (1998). Ethnopsychologies: Cultural Variations in Theories of Mind. page 30.

Liu, D., a Sabbagh, M., Gehring, W. J., and Wellman, H. M. (2004). Decoupling beliefs from reality in the brain: An ERP study of theory of mind. *NeuroReport*, 15(6):991–995.

Lohmann, H. and Tomasello, M. (2003). The Role of Language in the Development of False Belief Understanding: A Training Study. *Child Development*, 74(4):1130–1144.

Loukusa, S., Leinonen, E., Kuusikko, S., Jussila, K., Mattila, M.-L., Ryder, N., Ebeling, H., and Moilanen, I. (2007). Use of Context in Pragmatic Language Comprehension by Children with Asperger Syndrome or High-Functioning Autism. *J Autism Dev Disord*, 37(6):1049–1059.

Luo, Y. and Beck, W. (2010). Do you see what I see? Infants' reasoning about others' incomplete perceptions. *Developmental science*, 13(1):134–42.

Luyster, R. and Lord, C. (2009). Word Learning in Children with Autism Spectrum Disorders. *Developmental Psychology*, 45(6):1774–1786.

Lyn, H., Russell, J. L., and Hopkins, W. D. (2010). The impact of environment on the comprehension of declarative communication in apes. *Psychological Science*, 21(3):360–365.

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk.The CHILDES Project: Tools for Analyzing Talk. Third Edition*. Mahwah, NJ: Lawrence Erlbaum Associates, 3 edition.

Malle, B. F. (2002). The relation between language and theory of mind in development and evolution. In Givón, T. and Malle, B. F., editors, *The Evolution of Language out of Pre-Language*, pages 265–284. Amsterdam: John Benjamins.

Markman, E. M. (1990). Constraints Children Place on Word Meanings. *Cognitive Science*, 14(1):57–77.

Markman, E. M. and Wachtel, G. F. (1988). Children's Use of Mutual Exclusivity to Constrain the Meanings of Words. *Cognitive Psychology*, 20:121–157.

Martin, A. and Santos, L. R. (2016). What Cognitive Representations Support Primate Theory of Mind? *Trends in Cognitive Sciences*, 20(5):375–382.

Martin, I. and McDonald, S. (2003). Weak coherence, no theory of mind, or executive dysfunction? Solving the puzzle of pragmatic language disorders. *Brain and Language*, 85(3):451–466.

Mayer, A. and Trauble, B. E. (2013). Synchrony in the onset of mental state understanding across cultures? A study among children in Samoa. *International Journal of Behavioral Development*, 37(1):21–28.

Maynard Smith, J. and Harper, D. G. C. (1995). Animal signals: Models and terminology. *Journal of Theoretical Biology*, 177(3):305–311.

Maynard Smith, J. and Price, G. R. (1973). The logic of animal conflict. *Nature*, 246(5427):15–18.

Meins, E., Fernyhough, C., Wainwright, R., Das Gupta, M., Fradley, E., and Tuckey, M. (2002). Maternal Mind-Mindedness and Attachment Security as Predictors of Theory of Mind Understanding. *Child Development*, 73(6):1715–1726.

Meristo, M., Hjelmquist, E., and Morgan, G. (2011). How access to language affects theory of mind in deaf children. In Siegal, M. and Surian, L., editors, *Access to Language and Cognitive Development*, pages 44–61. Oxford University Press.

Meristo, M., Morgan, G., Geraci, A., Iozzi, L., Hjelmquist, E., Surian, L., and Siegal, M. (2012). Belief attribution in deaf and hearing infants. *Developmental Science*, 15(5):633–640.

Meristo, M., Strid, K., and Hjelmquist, E. (2016). Early conversational environment enables spontaneous belief attribution in deaf children. *Cognition*, 157:139–145.

Milligan, K., Astington, J. W., and Dack, L. A. (2007). Language and theory of mind: Meta-analysis of the relation between language ability and false-belief understanding. *Child development*, 78(2):622–46.

Moeller, M. P. and Schick, B. (2006). Relations Between Maternal Input and Theory of Mind Understanding in Deaf Children. *Child Development*, 77(3):751–766.

Moore, C. and Corkum, V. (1994). Social understanding at the end of the first year of life. *Developmental Review*, 14(4):349–372.

Moore, R. (2014). Ontogenetic Constraints on Grice's Theory of Communication. In Matthews, D., editor, *Pragmatic Development in First Language Acquisition*, pages 87–104. London: John Benjamins Publishing.

Moore, R. (2016a). Gricean Communication and Cognitive Development. *Philos Q*, 67(267):303–326.

Moore, R. (2016b). Gricean Communication, Joint Action, and the Evolution of Cooperation. *Topoi*.

Moore, R. (2016c). Meaning and ostension in great ape gestural communication. *Animal Cognition*.

Morgan, G., Meristo, M., Mann, W., Hjelmquist, E., Surian, L., and Siegal, M. (2014). Mental state language and quality of conversational experience in deaf and hearing children. *Cognitive Development*, 29:41–49.

Naigles, L. R. and Swensen, L. D. (2008). Syntactic Supports for Word Learning. In *Blackwell Handbook of Language Development*, pages 212–231. Wiley-Blackwell.

Nappa, R., Wessel, A., McEldoon, K. L., Gleitman, L. R., and Trueswell, J. C. (2009). Use of Speaker's Gaze and Syntax in Verb Learning. *Language Learning and Development*, 5(4):203–234.

Navarro, D. J., Perfors, A., Kary, A., Brown, S. D., and Donkin, C. (2018). When Extremists Win: Cultural Transmission Via Iterated Learning When Populations Are Heterogeneous. *Cognitive Science*, 42(7):2108–2149.

Novogrodsky, R. (2013). Subject pronoun use by children with autism spectrum disorders (ASD). *Clinical Linguistics & Phonetics*, 27(2):85–93.

Nunn, C. L. (2011). *The Comparative Approach in Evolutionary Anthropology and Biology*. University of Chicago Press.

O'Grady, C., Kliesch, C., Smith, K., and Scott-Phillips, T. C. (2015). The ease and extent of recursive mindreading, across implicit and explicit tasks. *Evolution and Human Behavior*, pages 1–10.

Olson, D. R. (1988). On the origins of beliefs and other intentional states in children. In *Developing Theories of Mind*, pages 414–426. Cambridge University Press, New York, NY, US.

Onishi, K. H. and Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science*, 308(5719):255–8.

O'Reilly, K., Peterson, C. C., and Wellman, H. M. (2014). Sarcasm and advanced theory of mind understanding in children and adults with prelingual deafness. *Dev Psychol*, 50(7):1862–1877.

Ozonoff, S., Cook, I., Coon, H., Dawson, G., Joseph, R. M., Klin, A., McMahon, W. M., Minshew, N., Munson, J. A., Pennington, B. F., Rogers, S. J., Spence, M. A., Tager-Flusberg, H., Volkmar, F. R., and Wrathall, D. (2004). Performance on Cambridge Neuropsychological Test Automated Battery subtests sensitive to frontal lobe function in people with autistic disorder: Evidence from the Collaborative Programs of Excellence in Autism network. *J Autism Dev Disord*, 34(2):139–150.

Parish-Morris, J., Hennon, E. A., Hirsh-Pasek, K., Golinkoff, R. M., and Tager-Flusberg, H. (2007). Children With Autism Illuminate the Role of Social Intention in Word Learning. *Child Development*, 78(4):1265–1287.

Paulus, M. and Fikkert, P. (2014). Conflicting Social Cues: Fourteen- and 24-Month-Old Infants' Reliance on Gaze and Pointing Cues in Word Learning. *Journal of Cognition and Development*, 15(1):43–59.

Penn, D. C. and Povinelli, D. J. (2007). On the lack of evidence that non-human animals possess anything remotely resembling a 'theory of mind'. *Philosophical transactions of the Royal Society B*, 362(1480):731–744.

Pennington, B. F. and Ozonoff, S. (1996). Executive functions and developmental psychopathology. *J Child Psychol Psychiatry*, 37(1):51–87.

Perez-Zapata, D., Slaughter, V., and Henry, J. D. (2016). Cultural effects on mindreading. *Cognition*, 146:410–414.

Perner, J. and Wimmer, H. (1985). "John thinks that Mary thinks that..." attribution of second-order beliefs by 5- to 10-year-old children. *Journal of Experimental Child Psychology*, 39(3):437–471.

Peterson, C. C. and Siegal, M. (1995). Deafness, Conversation and Theory of Mind. *Journal of Child Psychology and Psychiatry*, 36(3):459–474.

Peterson, C. C. and Siegal, M. (2000). Insights into Theory of Mind from Deafness and Autism. *Mind and Language*, 15(1):123–145.

Peterson, C. C., Wellman, H. M., and Slaughter, V. (2012). The Mind Behind the Message: Advancing Theory-of-Mind Scales for Typically Developing Children, and Those With Deafness, Autism, or Asperger Syndrome. *Child Development*, 83(2):469–485.

Piantadosi, S. T., Tily, H., and Gibson, E. (2012). The communicative function of ambiguity in language. *Cognition*, 122(3):280–291.

Pickles, A., Anderson, D. K., and Lord, C. (2014). Heterogeneity and plasticity in the development of language: A 17-year follow-up of children referred early for possible autism. *Journal of Child Psychology and Psychiatry*, 55(12):1354–1362.

Povinelli, D. J., Theall, L. A., Reaux, J. E., and Dunphy-Lelii, S. (2003). Chimpanzees spontaneously alter the location of their gestures to match the attentional orientation of others. *Animal Behaviour*, 66:71–79.

Powell, L. J., Hobbs, K., Bardis, A., Carey, S., and Saxe, R. (2018). Replications of implicit theory of mind tasks with varying representational demands. *Cognitive Development*, 46:40–50.

Preissler, M. A. and Carey, S. (2005). The role of inferences about referential intent in word learning: Evidence from autism. *Cognition*, 97(1):B13–B23.

Pyers, J. E. and de Villiers, P. A. (2013). Theory of mind in deaf children: Illuminating the relative roles of language and executive functioning in the development of social cognition. In Baron-Cohen, S., Tager-Flusberg, H., and Lombardo, M., editors, *Understanding Other Minds: Perspectives from Developmental Social Neuroscience*. Oxford University Press, Oxford, third edition edition. OCLC: ocn832601523.

Pyers, J. E. and Senghas, A. (2009). Language Promotes False-Belief Understanding: Evidence From Learners of a New Sign Language. *Psychological Science*, 20(7):805–812.

Quine, W. V. O. (1960). Chap 1: Language and Truth. In *Word and Object*.

Rakoczy, H. (2012). Do infants have a theory of mind? *The British Journal of Developmental Psychology*, 30:59–74.

Rubio-Fernández, P. and Geurts, B. (2012). How to Pass the False-Belief Task Before Your Fourth Birthday. *Psychological Science*, 24(1):27–33.

Russell, J. L., Lyn, H., Schaeffer, J. A., and Hopkins, W. D. (2011). The role of socio-communicative rearing environments in the development of social and physical cognition in apes. *Developmental Science*, 14(6):1459–1470.

Scarantino, A. (2013). Rethinking functional reference. *Philosophy of Science*, 80(5):1006–1018.

Schel, A. M., Machanda, Z., Townsend, S. W., Zuberbühler, K., and Slocombe, K. E. (2013a). Chimpanzee food calls are directed at specific individuals. *Animal Behaviour*, pages 1–11.

Schel, A. M., Townsend, S. W., Machanda, Z., Zuberbühler, K., and Slocombe, K. E. (2013b). Chimpanzee Alarm Call Production Meets Key Criteria for Intentionality. *PLOS ONE*, 8(10):1–11.

Schick, B., de Villiers, P., de Villiers, J., and Hoffmeister, R. (2007). Language and Theory of Mind: A Study of Deaf Children. *Child Development*, 78(2):376–96.

Schuwerk, T., Priewasser, B., Sodian, B., and Perner, J. (2018). The robustness and generalizability of findings on spontaneous false belief sensitivity: A replication attempt. *Open Science*, 5(5):172273.

Scott-Philips, T. (2014). *Speaking Our Minds*. Palgrave Macmillan.

Scott-Phillips, T., Blancke, S., and Heintz, C. (2018). Four misunderstandings about cultural attraction. *Evolutionary Anthropology: Issues, News, and Reviews*, 27(4):162–173.

Scott-Phillips, T. C. (2015a). Meaning in animal and human communication. *Animal Cognition*, pages 801–805.

Scott-Phillips, T. C. (2015b). Nonhuman Primate Communication, Pragmatics, and the Origins of Language. *Current Anthropology*, 56(1):56–80.

Senju, A., Southgate, V., White, S., and Frith, U. (2009). Mindblind Eyes: An Absence of Spontaneous Theory of Mind in Asperger Syndrome. *Science*, 325(5942):883–885.

Seyfarth, R. M., Cheney, D. L., and Marler, P. (1980). Monkey Responses to Three Different Alarm Calls: Evidence of Predator Classification and Semantic Communication. *Science*, 210(4471):801–803.

Shahaeian, A., Nielsen, M., Peterson, C. C., and Slaughter, V. (2013). Cultural and Family Influences on Children's Theory of Mind Development: A Comparison of Australian and Iranian School-Age Children. *Journal of Cross-Cultural Psychology*, 45(4):555–568.

Shannon, C. E. (1948). A Mathematical Theory of Communication. *The Bell System Technical Journal*.

Siller, M. and Sigman, M. (2008). Modeling longitudinal change in the language abilities of children with autism: Parent behaviors and child characteristics as predictors of change. *Developmental Psychology*, 44(6):1691–1704.

Siskind, J. M. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, 61(1-2):39–91.

Skyrms, B. (2004). *The Stag Hunt and the Evolution of Social Structure*. Cambridge University Press.

Slaughter, V. and Perez-Zapata, D. (2014). Cultural Variations in the Development of Mind Reading. *Child Development Perspectives*, 8(4):237–241.

Slaughter, V., Peterson, C. C., and Mackintosh, E. (2007). Mind What Mother Says: Narrative Input and Theory of Mind in Typical Children and Those on the Autism Spectrum. *Child Development*, 78(3):839–858.

Slaughter, V. P. and Peterson, C. C. (2011). How conversational input shapes theory of mind development in infancy and early childhood. In Siegal, M. and Surian, L., editors, *Access to Language and Cognitive Development*, pages 3–22. Oxford University Press.

Slocombe, K. E. and Zuberbühler, K. (2007). Chimpanzees modify recruitment screams as a function of audience composition. *PNAS*, 104(43):17228–17233.

Smith, A. D. M. (2014). Models of language evolution and change. *Wiley Interdisciplinary Reviews: Cognitive Science*, 5(3):281–293.

Smith, K. (2009). Iterated learning in populations of Bayesian agents. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, pages 697–702.

Smith, K. (2018). How Culture and Biology Interact to Shape Language and the Language Faculty. *Topics in Cognitive Science*, 0(0).

Smith, K. and Kirby, S. (2008). Cultural evolution: Implications for understanding the human language faculty and its evolution. *Philosophical Transactions of the Royal Society B*, 363(1509):3591–603.

Smith, L. and Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, 106(3):1558–1568.

Smith, N. J., Goodman, N. D., and Frank, M. C. (2013). Learning and using language via recursive pragmatic reasoning about other agents. In Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 26 (NIPS 2013)*, pages 3039–3047.

Southgate, V., Chevallier, C., and Csibra, G. (2010). Seventeen-month-olds appeal to false beliefs to interpret others' referential communication. *Developmental Science*, 13(6):907–912.

Southgate, V., Senju, A., and Csibra, G. (2007). Action Anticipation Through of False Belief by Attribution. *Psychological Science*, 18(7):587–592.

Sperber, D. (1996). *Explaining Culture: A Naturalistic Approach.*

Sperber, D. (2000). Metarepresentations in an evolutioinary perspective. In Sperber, D., editor, *Metarepresentations: A Multidisciplinary Perspective.* Oxford: OUP.

Sperber, D. and Wilson, D. (1995). *Relevance: Communication and Cognition.* Blackwell Publishing, 2 edition.

Sperber, D. and Wilson, D. (2002). Pragmatics, Modularity and Mind-reading. *Mind & Language*, 17(1-2):3–23.

Steels, L. (1996). Emergent Adaptive Lexicons. In Maes, P., editor, *From Animals to Animats 4: Proceedings of the Fourth International Conference on Simulating Adaptive Behavior*, Cambridge. MA: MIT Press.

Sterelny, K. (2012). Language, gesture, skill: The co-evolutionary foundations of language. *Philosophical Transactions of the Royal Society B*, 367:2141–2151.

Stirling, L., Douglas, S., Leekam, S., and Carey, L. (2014). The use of narrative in studying communication in Autism Spectrum Disorders: A review of methodologies and findings. In Arciuli, J., editor, *Communication in Autism*, number 11 in Trends in Language Acquisition Research, pages 171–215. John Benjamins, Amsterdam.

Stout, D. (2011). Stone toolmaking and the evolution of human culture and cognition. *Philosophical Transactions of the Royal Society B*, 366(1567):1050–1059.

Summers, K. and Clough, M. E. (2001). The evolution of coloration and toxicity in the poison frog family (Dendrobatidae). *PNAS*, 98(11):6227–32.

Surian, L., Baron-Cohen, S., and Van der Lely, H. (1996). Are children with autism deaf to gricean maxims? *Cogn Neuropsychiatry*, 1(1):55–72.

Surian, L., Caldi, S., and Sperber, D. (2007). Attribution of beliefs by 13-month-old infants. *Psychological Science*, 18(7):580–6.

Tager-Flusberg, H. (1992). Autistic children's talk about psychological states: Deficits in the early acquisition of a theory of mind. *Child Dev*, 63(1):161–172.

Tager-Flusberg, H. and Joseph, R. M. (2005). How Language Facilitates the Acquisition of False-Belief Understanding in Children with Autism. In Astington, J. W. and Baird, J. A., editors, *Why Language Matters for Theory of Mind*. Oxford University Press.

Tager-Flusberg, H., Paul, R., and Lord, C. (2005). Language and Communication in Autism. In Volkmar, F. R., Paul, R., Klin, A., and Cohen, D., editors, *Handbook of Autism and Pervasive Developmental Disorders*, pages 335–364. John Wiley & Sons, Inc., Hoboken, NJ, USA.

Taumoepeau, M. and Ruffman, T. (2006). Mother and Infant Talk about Mental States Relates to Desire Language and Emotion Understanding. *Child Development*, 77(2):465–481.

Taumoepeau, M. and Ruffman, T. (2008). Stepping Stones to Others' Minds: Maternal Talk Relates to Child Mental Language and Emotion Understanding. *Child Development*, 79(2):284–302.

Thompson, B., Kirby, S., and Smith, K. (2016). Culture shapes the evolution of cognition. *PNAS*, 113(16):201523631.

Tomasello, M. (2000). The Social-Pragmatic Theory of Word Learning. *Pragmatics*, 10(4):401–413.

Tomasello, M. (2008). *Origins of Human Communication*. MIT Press.

Tomasello, M. and Barton, M. E. (1994). Learning words in nonostensive contexts. *Developmental Psychology*, 30(5):639–650.

Tomasello, M. and Carpenter, M. (2007). Shared intentionality. *Developmental Science*, 10(1):121–125.

Tomasello, M., Carpenter, M., Call, J., Behne, T., and Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences*, 28:675–735.

Tomasello, M., Melis, A. P., Tennie, C., Wyman, E., and Herrmann, E. (2012). Two Key Steps in the Evolution of Human Cooperation. *Current Anthropology*, 53(6):673–692.

Tomasello, M., Strosberg, R., and Akhtar, N. (1996). Eighteen-month-old children learn words in non-ostensive contexts*. *Journal of Child Language*, 23(1):157–176.

Toth, K., Munson, J., N. Meltzoff, A., and Dawson, G. (2006). Early Predictors of Communication Development in Young Children with Autism Spectrum Disorder: Joint Attention, Imitation, and Toy Play. *Journal of Autism and Developmental Disorders*, 36(8):993–1005.

Vu, T. V., Finkenauer, C., Huizinga, M., Novin, S., and Krabbendam, L. (2017). Do individualism and collectivism on three levels (country, individual, and situation) influence theory-of-mind efficiency? A cross-country study. *PLoS ONE*, 12(8):1–20.

Wellman, H. M. (2014). *Making Minds: How Theory of Mind Develops*. Oxford University Press.

Wellman, H. M., Cross, D., and Watson, J. (2001). Meta-Analysis of Theory-of-Mind Development: The Truth about False Belief. *Child Development*, 72(3):655–684.

Wharton, T. (2003). Natural Pragmatics and Natural Codes. *Mind & Language*, 18(5):447–477.

Whiten, A. and Erdal, D. (2012). The human socio-cognitive niche and its evolutionary origins. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1599):2119–2129.

Wimmer, H. and Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13:103–128.

Wodka, E. L., Mathy, P., and Kalb, L. (2013). Predictors of Phrase and Fluent Speech in Children With Autism and Severe Language Delay. *Pediatrics*, 131(4):e1128–e1134.

Woensdregt, M. S., Kirby, S., Cummins, C., and Smith, K. (2016). Modelling the co-development of word learning and perspective-taking. In *Proceedings of the 38th Annual Meeting of the Cognitive Science Society*.

Wrangham, R. and Carmody, R. (2010). Human adaptation to the control of fire. *Evolutionary Anthropology*, 19(5):187–199.

Xia, H., Wu, N., and Su, Y. (2012). Investigating the Genetic Basis of Theory of Mind (ToM): The Role of Catechol-O-Methyltransferase (COMT) Gene Polymorphisms. *PLOS ONE*, 7(11).

Xu, F., Dewar, K., and Perfors, A. (2009). Induction, overhypotheses, and the shape bias: Some arguments and evidence for rational constructivism. In *The Origins of Object Knowledge*, pages 263–284.

Yu, C. and Ballard, D. H. (2007). A unified model of early word learning: Integrating statistical and social cues. *Neurocomputing*, 70(13-15):2149–2165.

Yu, C. and Smith, L. B. (2012). Embodied attention and word learning by toddlers. *Cognition*, 125(2):244–262.

Yurovsky, D. (2017). A communicative approach to early word learning. *New Ideas in Psychology*, 50:73–79.

Yurovsky, D. and Frank, M. C. (2017). Beyond naïve cue combination: Salience and social cues in early word learning. *Developmental Science*, 20(2):1–17.

Ziatas, K., Durkin, K., and Pratt, C. (1998). Belief term development in children with autism, Asperger syndrome, specific language impairment, and normal development: Links to theory of mind development. *J Child Psychol Psychiatry*, 39(5):755–763.

Zuberbühler, K. (2013). Acquired mirroring and intentional communication in primates. *Language and Cognition*, 5(2-3):133–143.

Zuidema, W. H. (2003). How the Poverty of the Stimulus Solves the Poverty of the Stimulus. In Becker, S., Thrun, S., and Obermayer, K., editors, *Advances in Neural Information Processing Systems 15*, pages 51–58. MIT Press.