U.PORTO
FLUP FACULDADE DE LETRAS
UNIVERSIDADE DO PORTO

Programa Doutoral em Filosofia

Meaningful play: Signaling games in light of later Wittgenstein

José Pedro Faria e Vasconcelos Correia

D

2019

José Pedro Correia

Meaningful play

Signaling games in light of later Wittgenstein

Tese realizada no âmbito do Programa Doutoral em Filosofia,
orientada pelo Professor Doutor João Alberto Cardoso Gomes Pinto

Faculdade de Letras da Universidade do Porto

Setembro de 2019

Meaningful play

Signaling games in light of later Wittgenstein

José Pedro Correia

Tese realizada no âmbito do Programa Doutoral em Filosofia,
orientada pelo Professor Doutor João Alberto Cardoso Gomes Pinto

Membros do Júri

Presidente:

Vogais:

À minha mãe

Contents

Declaração de honra

Declaro que a presente tese é de minha autoria e não foi utilizada previamente noutro curso ou unidade curricular, desta ou de outra instituição. As referências a outros autores (afirmações, ideias, pensamentos) respeitam escrupulosamente as regras da atribuição, e encontram-se devidamente indicadas no texto e nas referências bibliográficas, de acordo com as normas de referenciação. Tenho consciência de que a prática de plágio e auto-plágio constitui um ilícito académico.

Porto, 30 de Setembro de 2019

José Pedro Correia

Agradecimentos

Em primeiro lugar, gostaria de agradecer à professora Sofia Miguens, ao Mind, Language, and Action Group (MLAG), e ao Instituto de Filosofia (IF), por terem aceite acolher-me para levar a cabo este projeto. Agradeço também ao professor João Alberto Pinto por ter aceite orientar a minha tese e pelo apoio que deu durante a sua execução. Adicionalmente, estou imensamente agradecido à Fundação para a Ciência e Tecnologia pelo apoio financeiro (bolsa número SFRH/BD/100437/2014), sem o qual a elaboração deste trabalho não teria sido possível. I also thank Michael Franke for supporting my application to this PhD grant.

Durante o período de desenvolvimento desta tese, beneficiei de discussões estimulantes (mesmo que não necessariamente ligadas aos tópicos aqui abordados) com outros membros do IF. Em particular, gostaria de agradecer aos professores responsáveis pelas cadeiras que frequentei no ano curricular, que com os tópicos lecionados e as conversas aí geradas ajudaram de uma maneira ou de outra a cristalizar algumas ideias contidas nesta tese, incluindo a professora Maria João Couto, a professora Sofia Miguens, o professor João Alberto Pinto, e o professor Paulo Tunhas. Estou adicionalmente grato ao MLAG por ter podido participar na organização de atividades no contexto do grupo, o que enriqueceu a experiência, e em particular à Diana Couto com quem mais colaborei nessas iniciativas. Agradeço a todos os outros membros do MLAG, e aos restantes participantes, com quem tive o prazer de trocar ideias no contexto dessas atividades. When it comes more directly to the contents of this thesis, I will always be indebted to Michael Franke and Elliott Wagner for introducing me to signaling games in a very inspiring way, and additionally to Michael for very fruitful collaborations since then, one of which is included in this thesis. I would also like to thank Jiwon Kim, not only for some highly stimulating conversations about philosophy, but for also taking the time to review and give feedback on part of this thesis.

Agradeço também a todos os amigos e amigas que em algum momento nestes quatro anos acompanharam este processo, não só em convívios e atividades para além do doutoramento, mas também tolerando os eventuais desabafos ou até estimulando as divagações filosóficas. Esquivo-me de mencionar nomes com receio de me esquecer de alguém, mas vocês sabem quem são. This goes also to my friends abroad. Thank you for your companionship and support during this journey. Finalmente, gostaria de agradecer aos meus pais por apoiarem este desvio de percurso, aparentemente pouco sábio, com a confiança que sempre demonstrarem em mim e nas minhas escolhas.

Resumo

Esta tese explora as ligações entre a filosofia tardia de Ludwig Wittgenstein e método de modelação dos jogos de sinalização (*signaling games*). Em particular, foco-me nas *Investigações Filosóficas*, nomeadamente em quatro tópicos aí desenvolvidos: as observações de Wittgenstein sobre filosofia e método, a sua imagem da linguagem como uma prática dinâmica e heterogénea, as suas críticas a algumas variantes da conceção do significado como uma forma de correspondência, e as suas discussões sobre o seguimento de regras. Defendo que o método dos jogos de sinalização fornece-nos ferramentas que permitem um estudo do significado que está no geral bem alinhado com esses aspetos da filosofia tardia de Wittgenstein. Adicionalmente, a comparação revela trabalho na área dos jogos de sinalização que pode ser visto como filosoficamente problemático à luz dos comentários de Wittgenstein, e aponta para potenciais melhorias e direções para trabalho futuro.

Palavras-chave: filosofia da linguagem, Wittgenstein, jogos de sinalização

Abstract

This thesis explores the connections between Ludwig Wittgenstein's later philosophy and the modeling framework of signaling games. In particular, I focus on the *Philosophical Investigations* and on four topics within it: Wittgenstein's remarks on philosophy and method, his picture of language as a dynamic and heterogeneous practice, his criticism of some variants of conceptions of meaning as correspondence, and his discussions on rules and rule-following. I argue that the signaling games framework provides us tools that enable a study of meaning that is generally in line with these aspects of Wittgenstein's later philosophy. Exploring the connections additionally reveals where some work in the signaling games literature goes philosophically astray in light of Wittgenstein's remarks, and points to ideas for improvements and future work.

Keywords: philosophy of language, Wittgenstein, signaling games

List of Figures

List of Tables

Introduction

Language is a pervasive presence in the lives of human beings. So much so, that the ability to use language in the ways that we do is often considered one of the quintessential characteristics that differentiate us from other animals. Throughout the ages, many have attempted to develop a better understanding of how language works, from philosophers, to linguists, psychologists, anthropologists, neuroscientists, and researchers from many other fields. Despite language's ubiquity in our practices, and the great deal of attention it has received, there can be a feeling that progress in developing the theoretical tools to understand it better has been extremely slow, especially when compared with the study of other phenomena in natural science. Although we have confidence in our growing knowledge of quarks and atoms, molecules and chemical reactions, genes and cell structure, planets and galaxies, and so many other things, we seem to struggle to provide a satisfying theoretical characterization of a sentence as simple as "Hello!"

This is especially the case when it comes to the aspect of language that is the focus of this thesis: meaning. Efforts to get a better theoretical grip on meaning go back to philosophers in Ancient Greece, and have especially become an important part of philosophical practice since the so-called linguistic turn (Rorty, 1967). Despite all the work philosophers put into the topic, consensual progress is lacking (*e.g.* Putnam, 1970). Why is that? Perhaps the pervasiveness of language in our lives is the very thing that makes the study of it difficult; maybe we are too close to see the forest for the trees. Perhaps our close acquaintance with it creates ingrained preconceptions that get in the way of a deeper understanding. Perhaps it is a problem of self-reference—the fact that we need language to theorize about language—that ties us up in knots. Perhaps philosophers have been asking the wrong questions. Or perhaps it is simply that the phenomenon under study is just more broad and complex than it seems at first glance.

One philosopher that came to realize most of these issues is Ludwig Wittgenstein. His early work (*e.g.* 1922) is seen by most as outlining a systematic theory of meaning at the forefront of the linguistic turn. His later work (*e.g.* 1953), however, is mostly read as fundamentally undermining the picture of language that underlies his earlier theorizing efforts. Some read his later remarks not only as forming a critique of how philosophers typically go wrong when thinking about language, but also as presenting a case against the very possibility of one ever developing a systematic account of meaning. I believe the latter conclusion is too strong. A great deal of Wittgenstein's later criticism is targeted against the mainstream philosophical conceptions of language and meaning of his time. Those conceptions are easily tied to systematic approaches to meaning in general, but this connection is contingent. A criticism of the former need not necessarily imply, I believe, a full rejection of the possibility of the latter. But it is important to explore why.

In this thesis, I focus on Wittgenstein's *Philosophical Investigations* (1953; 2009)[1]. How I read it, the book attempts to undermine various intuitions that underlie, but are not exclusive to, a family of philosophical ideas about meaning. These intuitions include a craving for exactness and objectivity, the idea of meaning as something possessed and determined by linguistic expressions, a tendency to ignore the multifarious and dynamic ways in which we use language and overgeneralize from limited examples, among others. I see the epitome of the picture of meaning criticized by Wittgenstein in the analytic approach that uses logic as its main theoretical tool. This approach can be traced back to the work of Gottlob Frege and is currently embodied by contemporary formal semantics. Despite the heterogeneity of the field, I believe that objections to most (if not all) of its variants could be raised along the lines of at least one of the Wittgenstein's criticism of the aforementioned intuitions. To those that agree with this reading, the predominance of formal semantics as comprising the most systematic approaches to meaning in philosophy may make it seem that agreeing with Wittgenstein's later remarks on language implies giving up on systematicity. I believe that we can have our cake and eat it too.

The main objective of this thesis is to argue that there is at least one approach for studying meaning that is both systematic and in line with Wittgenstein's remarks in the *Philosophical Investigations*: signaling games. This framework, introduced by David Lewis (1969) and brought back to life after almost 30 years by Brian Skyrms (1996), provides game-theoretical tools to model linguistic interactions between agents. It fosters the use of mathematical modeling and computer simulations to study the dynamics of particular scenarios of language use. It can be used in a way that does not fall prey to the intuitions Wittgenstein warns us about, and is furthermore aligned with his alternative picture of language and meaning. This thesis is an exploration of the idea that the framework of signaling games points to a way that can embrace both the lessons of the *Philosophical Investigations* and systematicity in the study of meaning.

[1] I make a large number of references to the *Philosophical Investigations* (Wittgenstein, 1953). In order to reduce the dependence on a particular edition of the book, I mostly refrain from using a publication year and page numbers in citations, and instead cite numbered sections using the § symbol. In some occasions, the paragraph within a section is specified by using lowercase letters in the standard alphabetical order. For example, §64b refers to the *Philosophical Investigations*, section 64, second paragraph. Quotes are taken from the revised 4th edition by P. M. S. Hacker and Joachim Schulte (2009). When these are used, publication year and page numbers to this edition are provided in the citation.

To be clear, there are other families of models in the literature that address issues of communication in a similar way. One that is very closely related and used a lot in economics sprouted from a seminal paper by Michael Spence (1978). I will briefly mention some work belonging to this approach in this thesis, but the literature is much richer that what I can cover here. Connelly et al. (2011) give a more thorough overview. Another strongly related approach originates from the work of Luc Steels (1995) and is popular with researchers from the areas of statistical physics and complexity theory. An idea of the work developed by this lineage of research can be found, for example, in publications by Beckner et al. (2009) and Loreto, Baronchelli, and Puglisi (2010). The strong similarities between these approaches makes this thesis hopefully also relevant for these types of models. But, for the sake of brevity and clarity, I will mostly focus on the literature that stems from the work of Lewis (1969) and Skyrms (1996). This family of models comprises what in this thesis I call the *signaling games framework.*

I would like to also state what this thesis is not. First, it is not an in-depth historical exegesis of the *Philosophical Investigations.* I do not explore the history and origins of the text or make detailed contextualization of the ideas presented. My reading is mostly immanent (see Glock, 2007), taking the published text as a finished product, with only occasional references to *The Blue and Brown Books* (Wittgenstein, 1958, 2002) for additional clarification. Second, it is not an exploration of all topics presented in the book, but merely of those relevant to the use of the signaling games framework. This mostly includes issues relating to philosophical method, and both negative and positive remarks on language and meaning. Third, it is not a defense of the positions expressed in the book. The objective is not to supplement Wittgenstein's arguments with my own, but simply to provide an exposition of his ideas as they strike me. Naturally, this thesis is relevant insofar as one agrees with most of the remarks discussed here. It is unavoidable that my opinions additionally seep into the way I present Wittgenstein's ideas, but this thesis should not be seen as necessarily supporting those ideas.[2] Fourth, it is also not a full defense of the signaling games framework. Only arguments pertaining to Wittgenstein's remarks will be considered. This thesis is best thought of as attempting a defense of the following suggestion: if you like the reading of the *Philosophical Investigations* presented here, and are looking for a systematic approach to meaning that does not clash with it, you might be interested in looking into the signaling games framework.

The thesis is structured as follows. Part I consists of background information necessary to understand the comparison between Wittgenstein's ideas and the frame-

[2] If that was the case, much more argumentation would be required; the current thesis would be severely lacking.

work of signaling games. In Chapter 1, I discuss the issues of interpretation of Wittgenstein's work that were briefly mentioned in this introduction. I give an overview of the approaches in the literature and position my own reading in that context. An overview of the signaling games framework is provided in Chapter 2. This includes an introduction and a review of the literature and state of the art. The actual comparison begins in Part II. In Chapter 3, I consider Wittgenstein's remarks on methodology and their implications for the appropriateness of making use of the tools provided by the signaling games framework in the context of philosophy. In Chapter 4, I explore the picture of language set forth in the *Philosophical Investigations* and how well the signaling games approach embraces it. In Chapter 5, I address one of Wittgenstein's main target of criticism throughout the book: the intuition that linguistic expressions have meanings. I outline Wittgenstein's arguments against some variants of this idea, and explore their significance for certain conceptions of meaning within the signaling games literature. Finally, another important theme that is discussed at length in the *Philosophical Investigations* has to do with rules and rule-following. In Chapter 6, I explore Wittgenstein's remarks on this topic, and their relevance for possible interpretations of signaling game models.

This thesis additionally includes a multi-population signaling game model applied to vagueness. It was fully worked out during my doctoral program, and can be seen as an example of a signaling game model that reflects some of the suggestions stemming from taking Wittgenstein's remarks into account. The work itself (Correia and Franke, 2019) does not fit the main argument of this thesis, hence its inclusion here as an appendix (Appendix A). During the production of this thesis, another article closely related to it was published (Correia, 2019). In this case, the contents of the article already reflected the work that was being done for Sections 3.1 and 3.2, and some overlap exists with those sections.[3]

[3] Two more articles co-authored by me, based on previous work, were published during the course of the production of this thesis (Franke and Correia, 2018; Correia and Ocelák, 2019). These were not included here, since most of their contents were worked on prior to the start of my doctoral program.

Part I

Background

Chapter 1

Reading the *Philosophical Investigations*

> *Various interpretations of Wittgenstein and the* Philosophical Investigations *abound. In this chapter, I give an overview of some alternative ways of reading the book, and clarify my own personal approach.*

Interpretations of Ludwig Wittgenstein's work vary with respect to a number of different issues. There are differing opinions regarding whether one should take his writings as a unified vision or consisting of distinct phases, on how better to interpret his unusual writing style and rhetorical methods, on whether one can extract substantial views from his work and, if so, what exactly they are, among other things. Kahane, Kanterian, and Kuusela (2007, pp. 1-36) and Glock (2007) provide recent thorough overviews of the alternatives. His *Philosophical Investigations* (1953) is considered by many as a pivotal work in twentieth-century philosophy (Glock, 2007, p. 37), despite important disagreements on why that is so. Part of the reason for the various interpretations of this book in particular has to do with the writing style. Between different unidentified voices, rhetorical questions, meta-remarks on the discussion itself, it is often difficult to both clearly pin down which side of an issue is being argued for, and what the author's own opinion actually is (if there is any). It is, however, beyond the scope of this thesis to delve into this problem.[1]

Another reason for the heterogeneity of interpretations of the *Philosophical Investigations* stems from the interplay between how Wittgenstein conceives of philosophy, and the ways he actually goes about doing philosophy. A significant part of it (§§89-133) is dedicated to issues relating to philosophical method, with other remarks of the same nature scattered throughout the rest of the book. These are connected to the remaining remarks in a mutual relation. What Wittgenstein says

[1] See Stern (2004, pp. 21-28) for a more in-depth discussion.

about philosophy—what it should do and what it can achieve—should, one expects, have a direct impact on what he does when discussing certain philosophical topics, such as meaning, action, or the mind. But what is said about language in particular is also used to ground the metaphilosophical remarks. This can potentially be seen as a vicious cycle. Furthermore, there can seem to be, *prima facie*, an inconsistency between the two. Wittgenstein can appear at times to bluntly reject any kind of philosophical theorizing or even just the search for explanation, but he also sometimes seem to dogmatically defend particular views on some philosophical topics. In order to resolve this apparent tension, most interpreters argue that what he is saying about philosophy should make us see those apparent views as something different. Therefore, one can adumbrate a categorization of interpreters of Wittgenstein along the lines of how they see the relation between Wittgenstein's metaphilosophy and practice in the *Philosophical Investigations*.

In Section 1.1, I use this issue as a guide to map out some of the approaches in the literature. I roughly follow the overview by Kahane, Kanterian, and Kuusela (2007, pp. 1-36), and restrict myself to three main lines. In Section 1.2, I focus on another possible interpretation that, although often neglected, appeals strongly to me. In Section 1.3, based on this approach, I lay out my own choices of interpretation in order to make clear the type of reading of the *Philosophical Investigations* that can be expected in the remainder of this thesis.

1.1 Varieties of interpretation

Perhaps the most well-known reading of the *Philosophical Investigations* is the often called *orthodox* interpretation, epitomized by a series of four volumes by Gordon Baker and Peter Hacker (Baker and Hacker, 1980, 1985; Hacker, 1990, 1996), and subscribed to by many others. According to this reading, Wittgenstein's later philosophy presents a conception of meaning in opposition to a picture that underlies many approaches in philosophy of language: the idea of meaning as a form of correspondence between linguistic expressions (*e.g.* words, sentences) and other entities (*e.g.* objects in the world, ideas in the mind, senses belonging to a third realm[2]). The work of Gottlob Frege, Bertrand Russell, and Wittgenstein's own earlier work, the *Tractatus Logico-Philosophicus* (1922), are especially targeted.

Advocates of this interpretation further suggest that, in alternative to this picture, the key to a proper investigation of meaning lies, according to Wittgenstein, in the concepts of explanation and understanding:

[2] See Dummett (1996).

> The proper strategy is to focus on explanation and understanding and on the relation between them. In particular, giving a correct explanation is a criterion of understanding, while the explanation given is a standard for the correct use of the expression explained. Correspondingly, using an expression in accordance with correct explanations of it is a criterion of understanding, while understanding an expression presupposes the ability to explain it. (Baker and Hacker, 1980, p. 350)

The orthodox interpreter sees Wittgenstein as rejecting a particular idea about meaning and as grounding his practice on an alternative: language as rule-governed and grammar as a description of the rules of use. His metaphilosophy is said to derive from this picture of language and to be consistent with his practice. Wittgenstein is seen as both dispelling confusions originating from an incorrect picture of language, and providing clarifications by giving "an *Übersicht* of parts of the grammar of language that give rise to puzzlement" (Baker and Hacker, 1980, p. 368). The second aspect of Wittgenstein's practice should, however, not be misinterpreted as defending theories or theses, but rather as stating grammatical propositions, *i.e.* "familiar rules (grammatical rules) for the uses of words" (Hacker, 2012, p. 4). According to this view, the statement that "for the most part, the meaning of a word is its use" is no more a theory or a hypothesis than, for example, the statement that "red is darker than pink" (Hacker, 2012, p. 16).

Gordon Baker, although involved in the production of the first two of the aforementioned four volumes widely considered as the paradigm of the so-called orthodox interpretation, did not contribute to the latter two volumes because of a change of opinion regarding Wittgenstein's metaphilosophy (see Hacker, 2007). In his later work (Baker, 2004), he strongly rejects the idea that Wittgenstein was involved in any task of grammatical clarification, and defends instead the reading of the *Philosophical Investigations* as a book with a purely therapeutic aim. Baker specifically draws a strong link with psychoanalysis and argues that Wittgenstein's later work is aimed at releasing individuals from philosophical disquietude. According to this reading, Wittgenstein would be against presenting theses as philosophical assertions claiming truth and generality (Pichler, 2007, p. 126) and advancing whatever views that can appear as theses merely as highly problem- and person-specific alternative ways of looking at things. The particular analogy with psychoanalysis is present in other authors as well, one of the earliest proponents being John Wisdom (1953).

In its call for seeing Wittgenstein as making philosophy a purely therapeutic activity, dissolving problems with the objective of attaining philosophical quietude, this kind of reading has much in common with that of other authors who see Wittgenstein's later philosophy as a form of Pyrrhonian skepticism. One of the earliest advocates of this interpretation was Robert Fogelin (1976). Bob Plant (2004) gives a more recent proposal along similar lines. I will interchangeably call these approaches *therapeutic* or *quietist*. It is important to note, as John McDowell (2009) rightly points out, that Wittgensteinian quietism should not be equated with abandoning philosophy or gratuitously rejecting philosophical problems. It is an attempt at dissolving philosophical problems by actively engaging with them and showing that they were not interesting problems to begin with. This is not a passive task, as the label might suggest. The authors that defend these positions also claim consistency between Wittgenstein's metaphilosophy and practice: whatever apparent arguments or substantial views one finds in the book, should be seen as merely instrumental for achieving the therapeutic aim of attaining peace of mind.

Between the orthodox and the quietist interpretations, we thus find two different pictures of Wittgenstein's metaphilosophy. Morris (2007, pp. 74-76) sees this as a choice between attributing philosophy either two tasks or just one. In the orthodox interpretation, philosophy should have a negative task of dispelling confusions and misconceptions originating from a certain understanding of language and meaning. This would serve an instrumental purpose to make room for a positive task of providing the correct rules of use of the linguistic expressions involved in the misunderstandings. The negative task is thus, according to the authors sharing this reading of Wittgenstein, compatible with a positive task. In the therapeutic or quietist interpretations, however, there would be only one task for philosophy: undermining dogmatic ways of thinking. No substantial contribution should or could be made by philosophy since any proposal would reflect a single-minded view. Each of these readings of Wittgenstein's metaphilosophical remarks has implications for the reading of the rest of the book. Orthodox interpreters would endorse extracting substantial views from his work, at least in the sense of proposals about the correct use of linguistic expressions, while quietists would consider any apparent substantial views as merely alternative perspectives put forward to draw attention to the fact that there is never only one way of looking at things. They are pictures of how things might also be, rather than a defense of how things actually are.

The two readings have in common the assumption that Wittgenstein practiced what he preached. In order to defend the author's consistency of thought throughout the book, orthodox interpreters need to downplay the level of quietism in his metaphilosophical remarks, whereas quietists need to reduce the intended weight of Wittgenstein's apparent substantial views in the rest of the book. David Stern argues that one should not be too hasty in choosing one interpretation over the other:

> In order to understand the *Investigations*, we have to see that the tension between philosophy as therapy and philosophy as constructive argument operates there in a number of different ways. (2004, pp. 53-54)

Stern suggests that we see the book as reflecting the author's own vacillation between 'ending philosophy' and 'doing philosophy'. Although Wittgenstein was at times skeptic about what philosophy can or should do, he was also tempted to engage in philosophical argument.

Some evidence of this is circumstantial. Stern (2004, p. 53) quotes a passage where Rush Rees recounts a conversation with Wittgenstein in which the latter admitted that, even though he wrote that he could "break off philosophizing" (2009, §133c, p. 57e) when he wanted to, he actually found this difficult to do. But further evidence to support this view comes from Wittgenstein's own words in the preface to the *Philosophical Investigations*:

> After several unsuccessful attempts to weld my results together into such a whole, I realized that I should never succeed. The best that I could write would never be more than philosophical remarks; my thoughts soon grew feeble if I tried to force them along a single track against their natural inclination. — And this was, of course, connected with the very nature of the investigation. For it compels us to travel criss-cross in every direction over a wide field of thought. — The philosophical remarks in this book are, as it were, a number of sketches of landscapes which were made in the course of these long and meandering journeys. (2009, p. 3e)

If we take these words at face value, it seems difficult to see the book as putting forward a unified vision, be it about how language and meaning *actually* work, or about the limitations of the *whole* of philosophy. This might also help explain how easily proponents of each view are able to present a significant number of passages that support their interpretation and refute the others. Following Stern's approach, one would be unwise to try to take any apparent substantial views as forming a coherent and complete theory, but equally imprudent to dismiss them as merely instrumental to a global skepticism against philosophy.

Many authors in the *analytic* tradition take this realization as an avowal to draw philosophical theses or theories from passages in the book, despite Wittgenstein's metaphilosophical stance. This can be done in a vein that does not claim to be necessarily representing Wittgenstein's arguments accurately, but drawing on certain

interpretation	metaphilosophy	practice
orthodox	philosophy should provide an *Übersicht* of language	grammatical remarks
therapeutic/quietist	philosophy should exclusively dissolve problems	views as instrumental to therapy
analytic	separable from practice	skeptical arguments and/or proto-theories

Table 1.1: Rough categorization of different interpretations of Wittgenstein's metaphilosophy and practice in the *Philosophical Investigations*.

remarks to develop one's own perspective on the problem. The most well know example of taking this route is Saul Kripke's discussion of Wittgenstein's remarks on rule-following (1982). Kripke's approach to the metaphilosophical remarks in the the book is to simply brush them aside with a disclaimer. Other authors have taken different attitudes, like acknowledging them but willingly dissenting from them (*e.g.* Wright, 2007), downplaying their quietism in order to claim compatibility with their own analysis (*e.g.* Travis, 2006; Horwich, 2008). In all cases, though, the *Philosophical Investigations* is treated as a repository of either skeptical problems to be resolved in an analytic fashion or proto-theories to be further developed and detailed. The metaphilosophy is thus seen as not having a substantial bearing on the practice, allowing the analytic interpreter to dismiss it as a separate issue or mere idiosyncrasy.

Table 1.1 contains a rough summary of the distinctions introduced in this section regarding the varieties of interpretation of Wittgenstein's *Philosophical Investigations*. As any categorization, it is necessarily reductive and potentially paves over interesting nuances between interpreters. The objective is, however, to have an overview of some of the choices one has to consider when interpreting later Wittgenstein. The classification is also not exhaustive of all available options when it comes to the interpretation of the relation between Wittgenstein's metaphilosophy and practice. I mentioned David Stern's (2004) position as an introduction to the approach of the analytic interpreters, but Stern's reading is more nuanced. He considers the *Philosophical Investigations* as "a book that has a profoundly dialogical character" (2004, p. 37). Bob Plant, an advocate of a Pyrrhonian interpretation, does not think that all of Wittgenstein's views can be dismissed as merely instrumental, but that there is a set of ideas that form a core of minimal dogmatism (Plant, 2004, p. 242). And these are just examples of how each author's reading can vary.

In general, I agree with Stern that the book should not be coerced into a fully consistent whole, as both orthodox and therapist interpreters try to do. Wittgenstein says as much in the preface, and I don't see a reason to ignore his own words and try to project a unity that he himself did not see there. But, one should not simply brush aside the metaphilosophical position and draw on the rest of the material as a repository of views on traditional philosophical problems, like the analytic interpreters do. It is important to do "justice to the way in which these apparently incompatible aspects are intertwined" (Stern, 2004, p. 37) and keep both dimensions of Wittgenstein's thought in mind. Even though there might not be complete consistency, there is a certain harmony between the two aspects that some details in interpretation can bring to light. In the following section, I discuss another reading that I believe does this quite well: Richard Rorty's portrayal of later Wittgenstein as an edifying philosopher of the pragmatist kind.

1.2 Wittgenstein, the pragmatist?

In his defense of a metaphilosophical position commonly known as neopragmatism[3], Richard Rorty highlights a number of aspects of pragmatism in Wittgenstein's later philosophy. He is not alone in seeing this (*e.g.* R. Haack, 1982; Putnam, 1994; Blackburn, 2006, pp. 129-136). In this thesis, I will discuss Wittgenstein's contextualism, methodological pluralism, and anti-foundationalism. Additionally, considerations on explanations in the *Philosophical Investigations* come interwoven with remarks that emphasize language as a social practice, and with calls for seeing meaning in the broader context of other practices that involve the use of words and sentences. This goes in tandem with a rejection of realist and representationalist conceptions of meaning that see it as a form of correspondence between linguistic entities and other things. Although Wittgenstein never identified himself as a pragmatist, it is difficult to ignore these quintessentially pragmatist aspects of his picture of language and philosophy. Rorty's reading is, however, typically either ignored or classified as quietist (*e.g.* Glock, 2007, pp. 52-60). I believe that Rorty's reading is often misunderstood and is actually closer to the analytic interpretations than is typically acknowledged. There are, however, important nuances that set it apart from those, but these need to be understood in the context of his general outlook on philosophy.

[3] Although Rorty simply calls himself a pragmatist, other authors disagree (*e.g.* Putnam, 1994; S. Haack, 1997). Delving into this debate is beyond the scope of this thesis. I will use the term neopragmatism for Rorty's particular variety of pragmatism in order to signal the difference.

One important aspect of Rorty's neopragmatism is the criticism of a certain kind of philosophical attitude and practice. He has variously called it philosophy as pure subject or *Fach* (1976), systematic philosophy (1979), or simply Philosophy with capital P (1982). The target is a vision of philosophy as the study of certain paradigmatic problems that are somewhat exclusive to it. These have historically spanned over topics such as identity, subject and object, mind and matter, the nature and origin of knowledge, the relation between language, thought, and the world, among others. One belief characteristic of the attitude of Philosophy is that the problems it struggles with have definite solutions. The history of Philosophy is the story of the search for an Archimedean point of view from where all the apodictic truths about those issues would be clearly visible. From there, universal commensuration could be achieved, since the solutions to those problems would be beyond dispute. Rorty divides Philosophy into Platonists and Positivists, who both fit this description (1982, pp. xiii-xvii), differing only in whether they believe this belief is guaranteed by a transcendent realm, or by a correspondence between our claims and a mind-independent reality. According to Rorty, most ideas in the *Philosophical Investigations* can be seen as an attack on the mindset behind this vision of philosophical practice.

This picture of what Wittgenstein says philosophy should *not be*, is somewhat in line with the orthodox interpreters. The latter also argue that Wittgenstein defends there is nevertheless room for philosophy to positively contribute with grammatical clarifications on some of the issues Philosophy struggles with. Wittgensteinian quietists, on the other hand, think that there is no such constructive flip side to his philosophy. There are certainly a number of passages from Rorty's work that seem *prima facie* to support his classification in the latter category of interpreters. The following is a good example:

> In our time, Dewey, Wittgenstein, and Heidegger are the great edifying, peripheral, thinkers. All three make it as difficult as possible to take their thought as expressing views on traditional philosophical problems, or as making constructive proposals for philosophy as a cooperative and progressive discipline. (1979, p. 368)

Taken out of context, this seems like an adamant call for quietism. In my opinion, statements like this should be taken with a grain of salt, mainly because they were put forward in a context where Rorty was interested in emphasizing a contrast between so-called *edifying* philosophers and the approach of *systematic* philosophers.

In order to see this, it is crucial to note that Rorty is talking about 'views' as something tied to the latter:

> One has to think of philosophy as a name for the study of certain definite and permanent problems—deep-lying problems which any attempt at vision must confront: problems which professors of philosophy have a moral obligation to continue working on, whatever their current preoccupations. The Nature of Being, the Nature of Man, the Relation of Subject and Object, Language and Thought, Necessary Truth, the Freedom of the Will—this is the sort of thing which philosophers are supposed to have views about but which novelists and critics, historians and scientists, may be excused from discussing. It is such textbook problems which Wittgensteinians think the *Investigations* may let us dismiss. (1982, p. 31)

Rorty's adamant rejection of the idea that edifying philosophers 'hold views' does not therefore imply that he is saying that they are not putting forward any positive proposals. In his reading, Wittgenstein is someone that wants to abandon the ideas of truth as objective, language as in some special relation to reality, and philosophy as the rational quest for theories detailing those ideas. He is trying to undermine Philosophy by undermining its vocabulary (truth, language, proposition, etc) and replacing the way of thinking that is entangled with it, by changing the subject and talking instead about use, language-games, family resemblances, forms of life, and so forth. The former, rejecting a certain vocabulary, is a negative task, but the latter, proposing another, is a positive one. The difference between proposing a new vocabulary and defending substantial views, and the reason why Rorty rejects identifying one with the other, is that these positive proposals are of a different kind than those of Philosophy. Thus, we should "see edifying philosophers as conversational partners" (1979, p. 372) and, like them, set aside the idea that "when we say something we must necessarily be expressing a view about a subject" (1979, p. 371).

In order to establish the dichotomy between systematic and edifying philosophy, Rorty does emphasize Wittgenstein's negative task. His reading is, however, not a full quietist interpretation. In fact, the emphasizes of the positive task is more akin to the orthodox idea of grammatical clarification. But Rorty has something else in mind. Later, in a different context, and with perhaps a less strict conception of 'views', Rorty came to very explicitly endorse the picture of Wittgenstein as more than a pure quietist. In contrast with interpreters that have the latter picture he says that "[t]heir understanding of Wittgenstein's importance differs from that of philosophers who, as I do, find support in his writings for pragmatist views of truth and knowledge." (2007, p. 161) He calls the latter kind of interpreters "pragmatic Wittgensteinians" and goes on to characterize them as thinking that "their hero's importance consists in having replaced a bad theory about the relation between language and non-language, such as that offered in the Tractatus, with a better theory, the one offered in the Philosophical Investigations." (2007, p. 162)

Rorty's reading is thus that Wittgenstein's criticism of Philosophy is accompanied by a call for (and an exercise in) a kind of pragmatism. It is a vision of Wittgenstein as arguing against metaphysical essentialism, epistemological foundationalism, and linguistic representationalism, and proposing new ideas that do not presume any of the above. This is even more clear if, as argued by Kraugerud and Ramberg (2010), we see Wittgenstein as an example of Rorty's figure of the *ironist* (1989):

> Rorty's ironist holds that neither truth nor reality constrains human discursive practice; in other words, everything could be described differently than it has been. It is not that the ironist believes everything is up to the whim of human imagination. The ironist is simply not tempted to regard thought or language as (in any interesting sense) representing the world. (Kraugerud and Ramberg, 2010, p. 59)

The ironist recognizes that there is no Archimedean vantage point to underwrite our claims and beliefs, but does not succumb to a full-blown relativism. She recognizes the role of language as a coping tool and ventures to replace a certain vocabulary by a more useful one. The recognition that there is no absolute way to determine which vocabulary is the most useful—one true and final way of describing everything—does not keep the ironist from committing herself to certain vocabularies for certain tasks, and defend them against alternatives.

This relates again to the idea that there is a minimal dogmatism (Plant, 2004) in the *Philosophical Investigations.* Wittgenstein is indeed committed to certain ideas, in particular about language and meaning, that are presented in a somewhat dogmatic fashion. The orthodox interpretation attempts to give them an air of indisputability as mere descriptions of grammatical rules, but anyone that can imagine possible objections to them sees how that is just another form of Philosophical dogmatism. Reading Wittgenstein as a pragmatist ironist allows one to see these core views not as unjustified truth claims, but simply as the views Wittgenstein is committed to.[4] They, and the other views built upon them, are positive proposals, but

[4] Hacker (2012) makes a compelling argument that the orthodox position, with regards to this particular point, might be closer to Wittgenstein's original intentions than the interpretation I am suggesting here. However, to say that there is nothing dogmatic "about the grammatical proposition that there is no such thing as a private ostensive definition, or that the meaning of a word is not the object it stands for, or that for the most part, the meaning of a word is its use" (2012, p. 16) is, I believe, problematic, and at the very least flies in the face of the history

they are not anchored in some objective notion of truth.

In summary, reading Wittgenstein as a kind of pragmatist means acknowledging his criticism of a certain kind of doing philosophy, but leaving room for a certain kind of positive proposals. The distinction with the analytic interpretations is in the nature of these proposals. Under this interpretation, it is crucial to keep the following points in mind when considering apparent proposals in the *Philosophical Investigations.* First, proposals on a philosophical topic should not be seen as assertions about how things necessarily or objectively are, but rather as suggestions of alternative ways of thinking or talking about the topic. Second, positive proposals should not be taken as attempts to *fully* characterize a certain phenomenon (*e.g.* meaning and use), neither should negative arguments be seen as attempts to *completely* ban a certain vocabulary (*e.g.* meaning and reference). This point goes along with the idea that there are no hidden essences to be found. Third, arguments put forward to defend proposals, positive or negative, are not purported to be *the* ultimate unquestionable reasons why one should adopt those proposals. They can instead be seen as commitments of the author or means of convincing an audience.

1.3 A reading of later Wittgenstein

The interpretation adumbrated in the previous section squares well, I believe, not only with Wittgenstein's inclination to address typical philosophical questions and engage in argumentation, but also with his unorthodox rhetorical methods and the text's lack of formal structure. With its multiple voices, topic shifts, more and less focused parts, the text reads indeed like a conversation that reflects a pragmatist attitude towards doing philosophy: not forcing thoughts along a single

of philosophy. For each of those "grammatical propositions", one can find philosophers that have defended their opposite, even when well aware of those ideas. Even if we grant Wittgenstein that "not all grammatical propositions are immediately obvious" (Hacker, 2012, p. 14) and some require elucidation, it is difficult to see what the difference would then be between presenting a thesis and giving an argument for it, versus presenting a grammatical proposition and an elucidation of it. If that relies on the steps of the latter being "obvious and natural" (Hacker, 2012, p. 14), it shows a certain degree of dogmatism to assume that one's remarks can be supported by observations beyond dispute: either the *Philosophical Investigations* provides such elucidations, in which case it is impossible to explain why various readers of the book still dispute them, or it doesn't, in which case those remarks (or the steps of their elucidation) are as dogmatic as any other thesis in philosophy. One person's "obvious grammatical proposition" is another person's example of a dogmatic statement.

I leave open the possibility that Wittgenstein could have intended the position defended by Hacker (2012). This would, in my view, reveal a certain inconsistency with other aspects of his philosophy which, as I argued earlier, is definitely a possibility. The reading I am proposing here seems to me like a better way to square his metaphilosophy with some other views expressed in the *Philosophical Investigations* (namely those comprising his minimal dogmatism), but it is possible that it does not correspond to Wittgenstein's original intentions.

track but letting them "travel criss-cross in every direction" (2009, p. 3e). To try to coerce the *Philosophical Investigations* into a unified whole would be against Stern's advice, as discussed in section 1.1, of taking the dialogical character of the book seriously. The book is complex enough to contain elements of all that was brought forward so far—grammatical clarification, psychoanalytical therapy, Pyrrhonian skepticism—and potentially more. The reading approach delineated by Rorty seems consistent with Wittgenstein's metaphilosophical attitudes of anti-essentialism, anti-foundationalism, anti-representationalism, calls for abandoning the ideal of precision, and rejection of the metaphysical in favor of the everyday (more on this in Chapter 3). But is it the correct reading?

Reading the book from an ironist perspective additionally turns in on itself in an interesting way, as Alan Malachowski rightly points out:

> The 'reading' is more a case study, so to speak, in 'anti-essentialist' reading habits – that is, in pragmatism as applied to texts. For Rorty is suggesting that we will get more out of (say) *Philosophical Investigations* or *On Certainty* if we take it to be making fun of philosophers' theoretical pretensions. And this is an example of Rorty's view, amplified in his later work, that a 'good reading' of a text is *not* necessarily one that accurately extracts its 'intrinsic reading', its 'intended meaning' or any such thing, but one that puts the author's writing to the best extrinsic *practical use*. (2002, p. 95)

The approach is thus to go through a certain interpretation and see what one can get out of it in a given context. The ironist sees the writings of other philosophers as representative of particular vocabularies which one can explore by a process of redescription (Rorty, 1989, pp. 78-80). The aim is not to produce a 'true reading' (whatever that might be), but to undergo a hermeneutic experience. Interpretation is never a fully passive process: which aspect of the text will shine brighter depends on the perspective one takes and what one is interested in drawing from it. Rorty's advice is that one should simply acknowledge this and choose the interpretation that stimulates the most interesting developments, rather than squabble over whether it is or not correct. This seems to square well with the Wittgenstein's intention to "stimulate someone to thoughts of his own" (2009, p. 4e).

These suggestions typically raise fears that they warrant so-called interpretations that distort, misrepresent, or even completely disregard the original material. Although such exercises are certainly possible and not unheard of, it is not my intention to join those ranks. I aim to provide a reading of Wittgenstein's later philosophy that is not only coherent (with the aforementioned caveats in mind) but also plausible to others. With this in mind, I attempt to stay close to the text as much as possible in order for the reader to be able to track the source of my claims and judge their legitimacy themselves.

In the chapters to come, I will further guide my interpretation according to the following general principles. It will be a mostly immanent reading (see Glock, 2007, pp. 46-52), *i.e.* I will focus on the published text and its immediate context, rather than recurring to Wittgenstein's *Nachlass* or other unpublished material. I will consider Wittgenstein as at times proposing both positive and negative views along pragmatist lines, as delineated in section 1.2, while also in other moments advancing views for purely therapeutic purposes. I will reformulate and refer to such views in a manner close to Wittgenstein's: as remarks, rather than as theories or theses. The focus of this thesis being philosophical method, language, and meaning, I will restrict myself as much as possible to remarks on those topics.

Chapter 2

Signaling games*

In this chapter, I give a short introduction to the framework of signaling games, as well as an overview of some of the models in the literature.

Signaling games were created and introduced into philosophy by David Lewis (1969), later revived by Brian Skyrms (1996), and further explored by several other authors. The framework has also been used in economics, mainly stemming from the work of Michael Spence (1978), and theoretical biology, where it was imported by Alan Grafen (1990) and John Maynard Smith (1991). Because of their diverse origin and set of influences, terminology and notation is not always consistent throughout the literature.

David Lewis originally described signaling problems as involving a *communicator* and an *audience*. There are a number of alternative *states of affairs*, and the communicator knows which state holds. Based on this knowledge, the communicator can choose one of a number of alternative *signals*. The audience can tell which signal was chosen by the communicator and, based on this information, can choose one of a number of alternative *responses*. The behavior of the communicator can be represented as a mapping from states of affairs to signals, and this is called a *communicator's contingency plan*: it represents which signals the communicator chooses conditional on which state of affairs occurs. Similarly for the audience, one can represent its behavior (choices of responses conditional on signals chosen by the communicator) as a mapping from signals to responses, a so-called *audience's contingency plan*.

In most literature stemming from Lewis' work, however, this terminology is typically adapted to be more in line with the broader context of game theory. Since Brian Skyrms' revival of signaling games (1996) the most common nomenclature

*Some contents of this chapter (mostly Section 2.1) can also be found in the introduction to signaling games written for the publication entitled "Towards an Ecology of Vagueness" (Correia and Franke, 2019), included in this thesis as Appendix A.

is to call the communicator *sender* and the audience *receiver*; states of affairs are simply called *states*, signals are also referred to as *messages*, and responses are either called *acts* or *actions*; the set of all possible states (resp. messages, actions) is referred to as the state (resp. message, action) *space*; finally, contingency plans are called *strategies*, and so there is typically a *sender strategy* and a *receiver strategy.* Other alternatives are to call the sender *speaker*, the receiver *hearer* or *listener*, states *meanings*, and signals *forms.*

In the family of signaling models that stem from Michael Spence's work on job market signaling (1978) and are used mostly in economics and theoretical biology, rather than using the metaphor of the sender being aware of which state of affairs holds, authors talk about senders being of a certain *type*, where each type sends a particular signal. Types in this kind of models are formally equivalent to states in Lewis-Skyrms signaling games. In this thesis I give preference to Brian Skyrms' nomenclature, thus I will talk about *senders* and *receivers* having *strategies* of how to handle *states*, *messages*, and *actions.*

2.1 A short introduction

Lewis' original objective with the study of signaling games was to provide an answer to an argument raised by Quine (1936) and others against the possibility of language having started as a conventional system: if language is a convention, it had to be originally established by an agreement; in order to establish an agreement, a convention-governed system of communication would have to already have been in place; thus, although some languages could have been established by agreement if another convention was already in place, not all of them could. To this, Lewis retorts:

> I offer this rejoinder: an agreement sufficient to create a convention need not be a transaction involving language or any other conventional activity. All it takes is an exchange of manifestations of a propensity to conform to a regularity. (1969, pp. 87-88)

In order to support this claim, Lewis studies coordination problems formalized in terms of game theory. These are "situations of interdependent decision by two or more agents in which coincidence of interest predominates and in which there are two or more proper coordination equilibria" (1969, p. 24). In game theory terms, the agents interested in the coordination are the players, the game involves each player making an independent choice from his set of available choices, getting a payoff based on the choices of both.

Adapting one of Lewis' examples, say Alice and Bob want to get together. They usually meet at either Café One or Bistro Two. Imagine there is no way for them to make any explicit agreement about where to meet. They are thus left to independently decide to either go to Café One or go to Bistro Two and hope for the other to show up there. Neither has any preference for either place, but they do want to meet. Each thus prefers to go to one of the places only if the other also decides to go to that particular place. The setup of the coordination problem, *i.e.* the available choices and the relative interests of the players, can be represented in a payoff matrix, where rows show one player's available choices, columns the other player's available choices, and each cell gives the payoff for the players based on its row and column combination.

In the following, I assume Alice and Bob's interests are fully aligned, so we only need to specify one payoff value, which will be assigned to both. Consider the following matrix:

	b_1	b_2
a_1	1	0
a_2	0	1

We can see this as representing the following game: one player has a set of available choices $A = \{a_1, a_2\}$, the other $B = \{b_1, b_2\}$; they prefer to coordinate a_1 with b_1 or a_2 with b_2, thus if this is achieved each gets a payoff of 1, otherwise they each get 0. Connecting with the example laid out above, think of a_1 as Alice going to Café One, a_2 as Alice going to Bistro Two, b_1 as Bob going to Café One, and finally b_2 as Bob going to Bistro Two; their payoff is 1 for both if they coordinate on the place to meet (independently of which one), and 0 if they fail to do so. Formally, all we need to define in order to characterize such a kind of situation in general are the sets of choices A and B, and a utility function $U : A \times B \to \mathbb{R}$ that specifies the payoff for each player given the choices of both.

The above example is a very simple case, but it serves to illustrate Lewis' notion of *convention*. The two pairs of choices (a_1, b_1) and (a_2, b_2) are stable coordination equilibria, because in such a scenario no player has an incentive to unilaterally change his choice: if the first player is going to choose a_1, the second player would get a payoff of 0 for switching to b_2, instead of 1 from sticking with b_1; the same reasoning applies, *mutatis mutandis*, to the other player, and to both players in the other equilibrium. Such a combination of choices is a convention if, besides being a stable equilibrium, it is common knowledge that everyone conforms to those choices, and everyone expects everyone else to continue to conform to them. However, neither player has a preference between those two pairs. If the same coordination problem arises repeatedly, Lewis says, we can expect precedence to induce a kind of regularity: if the players manage to coordinate on one of the two equilibria, they should be expected (assuming they are rational and want to maximize their payoff) to repeat the choices that lead to that success, thus remaining in the same equilibrium. Their preference is to remain in a certain equilibrium given that others do too. This summarizes Lewis's general notion of a convention.

In order to extend this notion to *linguistic* conventions, Lewis considers situations where the choices available involve sending and receiving messages. Thus, we could think of two players with different roles. The sender has knowledge about which of a number of possible states obtains and, depending on this information, chooses a message to send. The receiver, on the other hand, has no direct knowledge about the state, but knows the message that the sender chose and, based on this information, chooses one of several possible actions. A preference relation exists between actions and states, and a payoff is attributed to each player based on the choices of both. Note that Lewis assumes that no player has any preference regarding the particular message that is used, provided that it enables coordination.

Formally, in order to describe the setup we need to specify a set of possible states T, a set of available messages M, a set of actions A, and the utility function $U : T \times A \to \mathbb{R}$. Despite the added dimension of the message exchange, these so-called signaling problems can be seen as particular cases of coordination problems if we consider the players' choices to be of whole strategies. A sender strategy is a specification of a choice of message for each possible state. It thus describes the sender's behavior conditional on the state that obtains. A receiver strategy analogously specifies a choice of action for each possible message. Thus, formally, what the sender chooses is a function $\sigma : T \to M$ and the receiver a function $\rho : M \to A$. Since agents choose strategies, we need a notion of utility that can be calculated in those terms. The expected utility (EU) of a pair of strategies, or strategy profile, (σ, ρ) can be calculated, using the utility function U, as a sum of the payoffs that would be obtained for all possible states, *i.e.*:

$$\mathrm{EU}(\sigma, \rho) = \sum_{t \in T} U(t, \rho(\sigma(t)))$$

As an example, consider a game with $T = \{t_1, t_2\}$, $M = \{m_1, m_2\}$, $A = \{a_1, a_2\}$, and the following utility matrix:

	$m_1 \mapsto a_1$	$m_1 \mapsto a_1$	$m_1 \mapsto a_2$	$m_1 \mapsto a_2$
	$m_2 \mapsto a_1$	$m_2 \mapsto a_2$	$m_2 \mapsto a_1$	$m_2 \mapsto a_2$
$t_1 \mapsto m_1, t_2 \mapsto m_1$	1	1	1	1
$t_1 \mapsto m_1, t_2 \mapsto m_2$	1	2	0	1
$t_1 \mapsto m_2, t_2 \mapsto m_1$	1	0	2	1
$t_1 \mapsto m_2, t_2 \mapsto m_2$	1	1	1	1

Table 2.1: Expected utility matrix.

	a_1	a_2
t_1	1	0
t_2	0	1

Since we have two states and two actions, this can be called a *binary signaling game*[2]. Possible sender and receiver strategies are, for example, $\sigma = \{t_1 \mapsto m_2, t_2 \mapsto m_1\}$ and $\rho = \{m_1 \mapsto a_2, m_2 \mapsto a_1\}$. These would have an expected utility of 2 for both sender and receiver, since when t_1 obtains the sender will use m_2 and to this message the receiver will respond with a_1 which achieves a payoff of 1, when t_2 obtains the sender will use m_1 and to this message the receiver will respond with a_2 which also achieves a payoff of 1. Calculations of expected utility unfold as follows:

$$\begin{aligned}
\mathrm{EU}(\sigma, \rho) &= \sum_{t \in T} U(t, \rho(\sigma(t))) \\
&= U(t_1, \rho(\sigma(t_1))) + U(t_2, \rho(\sigma(t_2))) \\
&= U(t_1, \rho(m_2)) + U(t_2, \rho(m_1)) \\
&= U(t_1, t_1) + U(t_2, t_1) \\
&= 1 + 1 \\
&= 2
\end{aligned}$$

Based on these calculations, one can consider all possible sender and receiver strategies and create a matrix of expected utilities. For this example, we obtain the matrix represented in Table 2.1. The two strategies just mentioned represent one of the two stable conventions in this game, the other being the pair of strategies $\sigma = \{t_1 \mapsto m_1, t_2 \mapsto m_2\}$ and $\rho = \{m_1 \mapsto a_1, m_2 \mapsto a_2\}$. Conventions of this kind in a signaling problem are what Lewis calls *signaling systems*. An example of complete miscoordination would be $\sigma = \{t_1 \mapsto m_1, t_2 \mapsto m_2\}$ and $\rho = \{m_1 \mapsto a_2, m_2 \mapsto a_1\}$. This is also called an anti-signaling profile. Partial coordination can be achieved, for example, by $\sigma = \{t_1 \mapsto m_1, t_2 \mapsto m_1\}$ and $\rho = \{m_1 \mapsto a_1, m_2 \mapsto a_1\}$. The latter are examples of what are called *pooling* strategies, *i.e.* strategies where either the sender

[2] This terminology is introduced by Hofbauer and Huttegger (2008).

$$\begin{tikzcd}
t_1 \arrow[r] & m_a \arrow[r] & a_1 \\
t_2 \arrow[r] & m_b \arrow[r] & a_2
\end{tikzcd}$$

(a) Signaling system

$$\begin{tikzcd}
t_1 \arrow[r] & m_a \arrow[dr] & a_1 \\
t_2 \arrow[r] & m_b \arrow[ur] & a_2
\end{tikzcd}$$

(b) Anti-signaling

$$\begin{tikzcd}
t_1 \arrow[r] & m_a \arrow[r] & a_1 \\
t_2 \arrow[ur] & m_b \arrow[ur] & a_2
\end{tikzcd}$$

(c) Pooling

Figure 2.1: Examples of strategy profiles.

(respectively receiver) does not differentiate between states (respectively messages) by always choosing the same message (respectively action). These example strategy profiles are illustrated visually in Figure 2.1.

Brian Skyrms (1996, pp. 80-104) identifies some problems with the story so far. Lewis' account of the stability of conventions rests on what could be considered strong demands on the agents. For there to be a certain degree of required common knowledge between them, which is necessary for a convention to hold, there needs to be a state of affairs that indicates to everyone involved that a certain regularity will hold, as well as "mutual ascription of some common inductive standards and background information, rationality, mutual ascription of rationality, and so on" (Lewis, 1969, pp. 56-57). These requirements can seem excessive, even more so if we consider how simple signaling systems are when compared to human languages. The models were introduced in order to help explain how language could get off the ground as a conventional system without any sort of prior agreement. However, if we consider the origins of language from a historical perspective, it seems implausible to assume a high degree of rationality of the agents that started making use of primordial signaling systems which (hypothetically) evolved into languages. Furthermore, communication through simple message exchange is something that almost all animals do: monkeys use calls, birds use singing, bees use dances, ants use pheromone trails, and so on. A plausible account of the origin of language should first explain how signaling systems like those could get started, without assuming a great deal of rationality from the part of the agents involved.

In order to address this problem, Skyrms proposes we study signaling problems in evolutionary terms. Rather than imagining, as Lewis does, rational agents making conscious decisions in possession of knowledge of the game and expectations of the behavior of other agents, we can imagine a simpler scenario inspired by biological evolution: there is a population of agents with biologically hardwired behaviors for engaging in interactions characteristic of a signaling problem; utility does not represent preference, but rather fitness for survival and reproduction; the make-up of the population evolves based on the relative fitness of the strategies present in the population. Such a setup attempts to capture the main features of natural selection: in a diverse population, agents with more successful strategies thrive, while

	$m_1 \mapsto a_1$ $m_2 \mapsto a_1$	$m_1 \mapsto a_1$ $m_2 \mapsto a_2$	$m_1 \mapsto a_2$ $m_2 \mapsto a_1$	$m_1 \mapsto a_2$ $m_2 \mapsto a_2$
$t_1 \mapsto m_1, t_2 \mapsto m_1$	s_1	s_2	s_3	s_4
$t_1 \mapsto m_1, t_2 \mapsto m_2$	s_5	s_6	s_7	s_8
$t_1 \mapsto m_2, t_2 \mapsto m_1$	s_9	s_{10}	s_{11}	s_{12}
$t_1 \mapsto m_2, t_2 \mapsto m_2$	s_{13}	s_{14}	s_{15}	s_{16}

Table 2.2: Possible strategy profiles in a binary signaling game.

agents with less fit strategies die off. Although the inspiration for this scenario is biological evolution, similar things could be said (*e.g.* Dawkins, 1976; Boyd and Richerson, 1985) about how ideas spread in a population of agents who can adopt or abandon them depending on how successful they prove to be. The principles can be captured in a formal model that abstracts away from the interpretations: the replicator dynamic. The only thing relevant to this equation are the relative proportions of strategies in a given population and the utility function. Using it, one can compute which strategies evolve under which conditions.

Let's see how this works for our ongoing example. Imagine a population composed of agents that play both as sender and as receiver. Their hardwired behavior therefore consists of a combination of two strategies, one for each role. We can represent each strategy profile as $s_i \in S : (T \to M) \times (M \to A)$, referring to its sender strategy as s_i^σ and its receiver strategy as s_i^ρ. There are 16 possible strategy profiles, and I will number them as shown in Table 2.2. If we represent the proportion of strategy profile s_i at time instant t as $x_i(t)$, we can define the step-wise changes in proportion in terms of the following equation:

$$x_i(t+1) = x_i(t)\frac{f_i(t)}{\bar{f}(t)}$$

In this equation, f_i represents the fitness of strategy profile s_i, and $\bar{f}$ the average fitness of all strategies in the population. The proportion of a strategy profile at time $t+1$ is thus dependent on its proportion and relative fitness at time t. Fitness can be calculated as follows:

$$f_i = \sum_{s_j \in S} x_j \left(\mathrm{EU}(s_i^\sigma, s_j^\rho) + \mathrm{EU}(s_j^\sigma, s_i^\rho)\right)$$

Thus, the fitness of a strategy profile is given by the sum of the proportional expected utility of playing as sender and receiver against each strategy profile in the population.

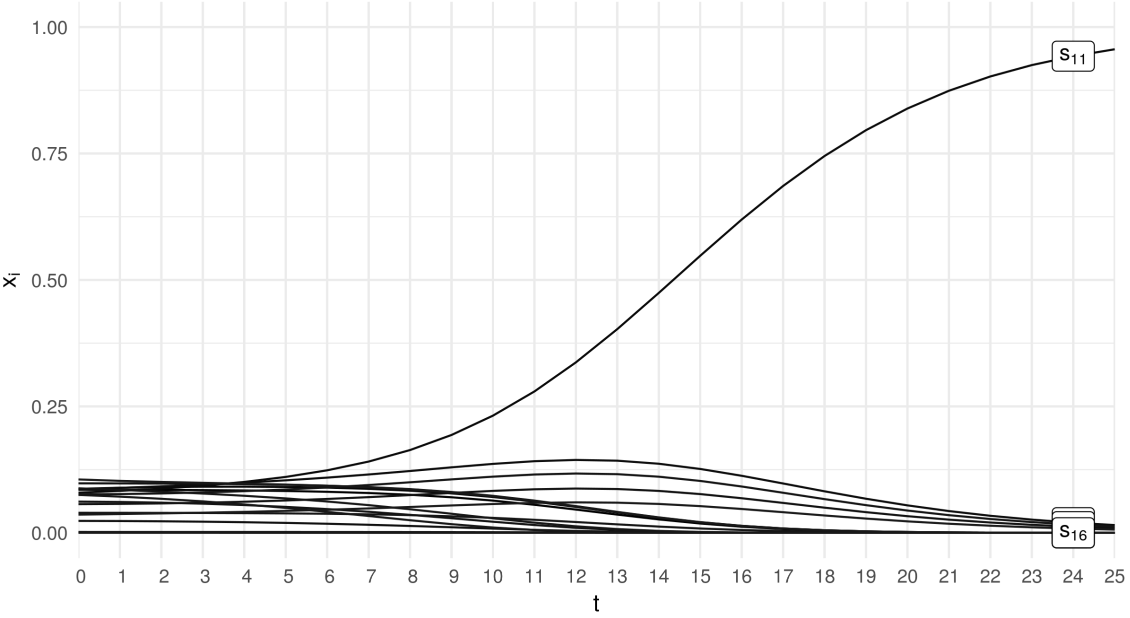
A simple way of seeing how strategy profiles fare against each other under this dynamic is to start with a population with random proportions and calculate their change over a number of time steps. Figure 2.2a shows the evolution of strategy profile proportions over time for one simulation run. Notice that the strategy profile that appears to be taking over the population (almost reaching proportion 1) is the signaling system s_{11} (see again Table 2.2). Since proportions add up to 1, we can represent several simulation runs in a ternary plot, like the one in Figure 2.2b, where the bottom left corner represents s_6 dominating the population, the bottom right corner represents s_{11} dominating the population, and the top corner represents any other strategy profile dominating. The plot shows how every simulation run was driving towards one of the signaling systems taking over the population.

The characterization of signaling problems in terms of evolutionary game theory allow us to explain why certain equilibria come to be and how. Not only can we better understand why signaling systems are stable even without any assumptions of rationality, but we can also map out which initial conditions drive the system towards which equilibria and which don't, as we have seen in the previous example. Although simulation results can only give an indication of this, it can be shown by other means that, for a simple binary signaling game like the example just explored, an evolutionary process of the kind described always drives the population into a state where one signaling system takes over completely (see Huttegger, 2007a). This is not always the case, though. More complex signaling problems may have different evolutionary outcomes, sometimes unexpected ones.

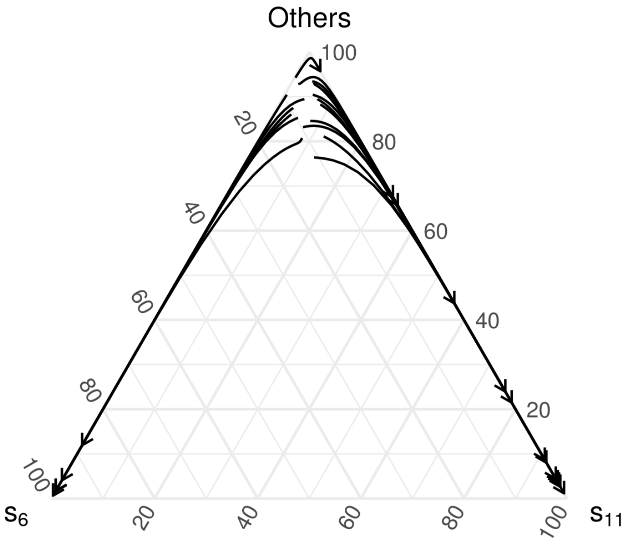
Skyrms (2010) gives an overview of different topics studied using signaling games, including expansions of the framework itself (for example, considering other dynamics beyond the replicator equation), exploration of other factors that impact the evolution of signaling (for example, how agents are interconnected), or variations on the signaling problem and its basic assumptions (for example, loosening the alignment of interests in order to provide accounts of deceptive signal use). Applications of signaling games include discussions of categorization, compositionality, incommensurability, and vagueness, just to name a few. More recent overviews are given by Huttegger (2014) and Huttegger, Skyrms, Tarrès, et al. (2014). In the following section, I give my own short overview by exploring the various ways in which the typical components of a signaling game model have been changed in the literature.

2.2 Anatomy of a signaling game

In this thesis, I talk about the signaling games approach not as a theory of meaning but rather as a framework. This distinction aims to capture a looser conceptual orientation. What the approach provides is a number of ingredients that can be



(a) Strategy proportions over time in one simulation run.



(b) Ternary plot of 20 simulation runs.

Figure 2.2: Simulation results of a binary signaling game under the one-population replicator dynamic. Plots created using the R libraries `ggplot2` (Wickham, 2016) and `ggtern` (N. E. Hamilton and Ferry, 2018).

combined to create models of communicative interactions. These include agents, states, messages, actions, and more. Each of these ingredients can come in different forms and variants are not limited *a priori.* The signaling games literature also contains various ways of studying the created models, from static analysis to computer simulations.

The approach does not aim to be complete when it comes to the kinds of interactions that can potentially be modeled. When a researcher is interested in exploring a particular aspect of a particular type of language use, what the framework provides is a general way to build an abstract model, and a number of components to potentially combine. Different problems require different tools, and if the toolbox does not contain what one needs for a particular use, new tools can be created. Others can be and have been imported from the development of game theory for other purposes and other areas like, for example, evolutionary biology or economics. The application of signaling games to a range of different problems has, since its inception, given rise to an ever-growing variety of components.

In this section, I want to give an overview of the conceptual tools available, and in the process give a more detailed survey of the different models in the literature. As a basis for comparison, I will refer to models of the kind exemplified in the previous section as *simple Lewis signaling games.* These are models where states, messages, and actions all come from discrete finite spaces, there is one and only one correct action for each state, there is only one sender and one receiver, and there is full common interest between them, *i.e.* the utility function rewards the same state-action combinations for both. Henceforth, a ixjxk Lewis game is a simple Lewis signaling game with i states, j messages, and k actions. The simplest model (of any interest) is a 2x2x2 Lewis game and can also be called a binary signaling game (Hofbauer and Huttegger, 2008).

States

In a Lewis game, states are simple elements of a discrete set. What this means is that states do not have any properties and are related to other states only by identity: the most that one can ask of two states is whether or not they are the same. Other state spaces are possible. First, one can imagine that not all states are equally likely to hold. This can be modeled by specifying a *prior probability* function that assigns to each state a value between 0 and 1 representing the probability of that state occurring. The default, with all states being equally likely, is equivalent to specifying a uniform prior probability function. This information can be incorporated and have an impact on the calculation of the expected utility of strategies and on how these strategies change in the context of a dynamic analysis.

Hofbauer and Huttegger (2008) study the impact of non-uniform priors in binary signaling games under two different dynamics. Their results show that the more skewed the prior distribution is, the less likely signaling systems will evolve. In certain conditions, some signaling strategies that constitute situations where no communication actually occurs (so-called pooling equilibria) even become optimal solutions to the game. A vivid example of the impact of different priors on the outcome of evolutionary processes can be found in the work of Jäger, Metzger, and Riedel (2011, p. 527). Priors can either be interpreted ontologically, *i.e.* as the real probabilities of states occurring, or epistemically, *i.e.* as the probability distribution that the agents project onto the state space (either implicitly given their behavior, or explicitly by holding different beliefs). In the latter case, it makes sense to model situations where priors by sender and receiver diverge. Brochhagen (2017) explores the consequences of this in the context of linguistic ambiguity and argues that having a vocabulary with some ambiguity can enable agents to better adapt to misaligned priors.

Another aspect where one can go beyond the classic Lewisian game is to add structure to the state space. One example is the family of models called *similarity maximization*, or sim-max for short, introduced by Jäger and van Rooij (2007). In these models, states are related to each other by a distance metric, effectively turning the state space into a metric space. This means that, besides asking whether or not two states are identical, one can additionally ask how distant they are from each other. The usual use is to define a similarity metric based on the distance, and to set up games where the task of the receiver is to guess the original state; the utility function is typically proportional to similarity, thus allowing the agents to be rewarded not only for guessing the exact state correctly, but also, to some extent, for getting close. These models have been used to study linguistic categorization (*e.g.* Jäger and van Rooij, 2007; Jäger, Metzger, and Riedel, 2011; O'Connor, 2014a, 2019; Correia and Ocelák, 2019), vagueness (*e.g.* Franke, Jäger, and van Rooij, 2011; Correia, 2013; O'Connor, 2014b; Franke and Correia, 2018; Correia and Franke, 2019)[3], and ambiguity (*e.g.* O'Connor, 2015).

Davis (2017) develops a sim-max model where each state is itself structured, being available to both sender and receiver as a vector rather than a simple element. This enables him to incorporate an additional step in the signaling model that aims to represent perception as the mapping of multi-dimensional values (*e.g.* an image as a vector of pixels) to lower-dimensional internal representations and study the

[3] The work by Correia and Franke (2019) is reproduced here in Appendix A.

effects of different perception models (*e.g.* Bayesian, artificial neural networks) on the evolution of associations between internal representations and signals. Finally, although the family of models stemming from Lewis and Skyrms typically uses discrete state spaces, Crawford and Sobel (1982), and the literature that builds upon their work, consider states as random variables drawn from a continuous distribution.

In the vast majority of signaling games studied in the literature, state spaces are additionally static: the set of possible states is defined by the modeler and remains the same throughout the analysis. Although variants of the same model with different state spaces are often considered (*e.g.* changing the number of states), each variant is studied independently. However, the world is not static and language needs to be able to adapt to a changing environment. Following this motivation, Alexander (2014) models two types of dynamic state spaces, one where new state-action pairs keep getting created, and another where the associations between states and actions keep getting changed. In this context, the author compares some learning models and finds that being able to discount the past is an important feature that allows reinforcement learning to successfully cope with a dynamic environment.

Messages

As with states, messages in signaling game models are typically simple elements of a discrete set, but other types of message spaces are possible. Nowak and Krakauer (1999) study state-matching games[4] where messages can be mistaken for one another. Much like state spaces in sim-max games, message spaces in their models consist of elements (the messages) and a similarity metric. This metric is then used to define the probability of one message being mistaken for another. In this noisy environment, the authors find that there is a limit to the number of message-state associations that possibly form under evolutionary dynamics. In the context of a large state space, this means that agents will evolve to ignore certain states, independently of how many messages they have available. This leads the authors to propose an additional way of making the message space more complex, namely by allowing different messages to be combined. In such a game, sender and receiver thus exchange a complex message composed of simpler ones, and this allows them to overcome the so-called linguistic error limit, *i.e.* “the number of objects that can be accurately described” (1999, p. 8029) in a protolanguage.

Complex messages can also be found in work related to the multiple-sender *syntactic games* of Barrett (2009). The underlying motivation behind the development of these games is to be able to model situations where messages are sent to a receiver

[4] See following section on actions.

from multiple sources. Each sender chooses one message independently from others, and the receiver makes a choice of act based on all messages. Thus, what the receiver handles is effectively a complex message composed of simpler ones. Another type of complex messages appear in the aforementioned work by Davis (2017) where, like states, they are also represented as multi-dimensional vectors. Hence each message can be seen as a complex combination of different features.

Both simple and complex messages still belong to discrete spaces, where the set of elements is predetermined and finite. In the family of models introduced by Spence (1978), however, messages are typically values of a continuous quantity. In his model, potential employees (senders) invest in education in order to signal their productivity type to the potential employer (receiver) who in turn decides the wage to assign them. Spence considers the signal to be the cost of the investment in education, which is modeled as a positive real number. This means that the messages space (set of possible messages to choose from) is both infinite and right-unbounded. The literature that makes use of this kind of models in economics is extensive, and an overview of its applications beyond the scope of this thesis. A thorough review is provided by Connelly et al. (2011).

All the message spaces discussed so far are static: the messages available to the agents are predetermined by the modeler and assumed not to change during the adaptive processes that shape the way sender and receiver use them. However, this need not be so. Alexander, Skyrms, and Zabell (2012) introduce a model[5] where new messages can be invented. For each state, the sender chooses either an existing message or a new one with a certain probability (which diminishes with the number of available messages). On getting a new message, the receiver chooses an action randomly (typically according to a uniform distribution over the action space). If the action turns out to be successful, both sender and receiver add the new message into the message space, otherwise they ignore it. Agents adapt their strategies using reinforcement learning, for both existing signals and successfully introduced new signals. The authors find that this dynamic prevents the evolution of sub-optimal strategies (pooling equilibria) that are increasingly likely to occur as the number of states or the inequality of their prior probabilities increases in models using a static message space. In order to avoid eventually ending up with a large number of fairly useless signals, the authors also incorporate a mechanism through which signals can be forgotten and removed from the message space. Just like the vocabulary of natural languages, message spaces can thus be modeled to be continuously open to change.

[5] Originally explored by Skyrms (2010, pp. 118-135).

Actions

Like states and messages, action spaces can be discrete or continuous, simple or structured, static or dynamic. The literature discussed so far covers most of these variants. States and actions are typically distinct entities, being related only through the utility function. A number of models, however, have the receiver choosing states rather than actions. This could be thought of as eliminating the action space from the model, or simply as equating it with the state space. The latter has the advantage of maintaining the general structure of a game.

The most simple example of such a model is a *state-matching* game, where agents are rewarded if and only if the receiver chooses the same state that the sender observes. One could imagine other possibilities, the most obvious being a kind of negation game with two states, where agents would be rewarded if and only if the receiver chooses the state *not* observed by the sender. Going a bit beyond this setup we have, for example, the aforementioned sim-max games of Jäger and van Rooij (2007), where agents are rewarded proportionally to how similar the receiver's choice is in comparison with the state observed by the sender.

These games are especially relevant when the number of states is much higher than the number of messages and agents cannot therefore develop a perfect state-action matching, having to make do with getting close enough. This applies to how we use a limited basic vocabulary to carve up more fine grained domains, such as for example when talking about people as 'tall' and 'short' in order to categorize them in terms of height, or when using basic color terms, such as 'red', 'green', and so on, to refer to certain parts of the potentially infinite visible spectrum (see Correia and Ocelák, 2019 for an example of work addressing the latter situation).

Utility

In general, actions, states, and messages are related to each other via the utility function. Depending on the interpretation of the model, utility can be seen as capturing either agents' preferences or relative evolutionary advantages of strategies. In simple Lewis signaling games, one typically specifies only one utility function that represents the payoff obtained by both sender and receiver for each combination of state and action. This assumes that interests or advantages are both fully aligned and symmetrical, *i.e.* that both agents benefit (or fail to benefit) equally in each possible situation. This is what is called a *common interest* or *pure coordination* game. In general, however, signaling is unlikely to develop or be maintained in such ideal conditions. In order to break the symmetry and make the model more general, independent utility functions for sender and receiver can be specified. Lewis already allowed for this from the start, and even reflected on specific examples where there is some degree of conflict between agents (1969, pp. 15, 71, 95, 117). He did, however, confine his attention to "situations in which coincidence of interest predominates" (Lewis, 1969, p. 14). But successful communication can also emerge in situations of partial conflict.

Skyrms (2010, pp. 73-82) investigates what happens to signal use in the context of partial conflict of interests between the agents involved in a signaling game. In general, as pointed out by Zollman, Bergstrom, and Huttegger (2013, p. 1), "when interests are not aligned, signallers might be selected to manipulate signal receivers with misleading signals, and the signal receivers might evolve to disregard such communications." This manipulation from the part of the sender is what Skyrms calls *deception*, and occurs whenever a sender systematically uses messages that raise the probability of the receiver choosing actions that benefit the sender more than the actions that would have been chosen had the receiver known the state that actually obtains. It is interesting to note that the deceptive use of some signals by the sender are not necessarily fully damaging to the receiver, and might even form part of an equilibrium where partially successful communication can occur. Skyrms discusses two scenarios where, although messages encode partial information regarding the states the sender observes, and the receiver's payoff is lower than it would be if a full signaling system would be in place, the strategies are nevertheless stable and the overall payoff is higher than in a scenario of full break down in communication. Martínez (2019) further argues that there are other possible ways of interpreting these situations as cooperative rather than deceptive.

According to Wagner (2012), partial communication is even possible in situations of totally opposed interests. In a zero-sum signaling game, the sender only profits from the loss of the receiver and vice versa. This makes any honest communication intuitively undesirable, since if the sender would use messages in a way that would convey some information about the observed states to the receiver, the latter could use that information against the sender, and that is exactly what one finds in such a system in equilibrium. Wagner studies a zero-sum signaling game where the agents' strategies never actually converge to an equilibrium, but keep adapting to each other in a non-regular pattern. Because of that, the amount of information conveyed by each message as the strategies continuously change never actually drops down to zero, *i.e.* "partial information transfer can be sustained indefinitely in out-of-equilibrium play" (2012, p. 25). Godfrey-Smith and Martínez (2013) and Martínez and Godfrey-Smith (2016) further explore the relation between common interest and successful communication by exploring a large number of scenarios between the two extremes of zero-sum and fully aligned interests. They corroborate Wagner's results, and find that the more aligned the interests of sender and receiver are, the more likely they are to develop some degree of communication.

The payoff structure of a simple Lewis signaling game is influenced only by states and actions. Since messages do not directly impact utility this type of communication falls into the category of what in game theory is usually called *cheap talk*. Alternatively, in models of *costly signaling*, sending certain messages can incur a decrease in utility for the agents involved. A family of models that makes use of this stems from Michael Spence's work on job market signaling (1978). In the original model, there are two types of senders, and a receiver that, not knowing their type, has to assign a salary level based on a signal from the sender. Senders are individuals applying for a job, and the two types represent different productive capabilities. The receiver (*e.g.* an employer hiring) wishes to assign a higher salary to the more productive individuals, and a lower salary to the less productive. As before, honest signaling (where individuals would send a different signal depending on their type) is impossible in this base scenario, since less productive individuals have an incentive to send the same signal as the more productive individuals do, in order to attempt to receive a higher salary. If, however, signaling as being a more productive individual has a lower cost for individuals that are actually of that type, honest signaling becomes an equilibrium under certain parameter values (see Spence (2002) and Wagner (2013) for more recent perspectives on this).

The same realization has been made in theoretical biology, first informally by Zahavi (1975, 1977), and subsequently demonstrated in a signaling game model by Grafen (1990). Maynard Smith (1991) made a model of this type of interaction popular with his Sir Philip Sidney game. In this game, the sender is imagined as an individual potentially in need of a resource, and the receiver as another individual deciding whether or not to donate that resource to the sender. In such a model, it is observed that signals must be costly for them to be honest in a situation of pure conflict, but if both agents have a stake on the survival of the other (in particular, in the case of this model, by being related to each other), honest signaling can still be evolutionarily stable even without costs (see Maynard Smith, 1994 for more on this). Some more recent analyses of these games (Huttegger and Zollman, 2010; Zollman, Bergstrom, and Huttegger, 2013) show that there are also stable hybrid equilibria where low-quality or not-in-need individuals sometimes signal honestly, but sometimes also deceive, and receivers act accordingly by not always taking a signal as honest.

Pacheco et al. (2015) show another application of costly signaling in theoretical biology. The authors model a collective action problem, where a population can produce a public good (shared equally by all) if a certain minimum number of individuals (though not necessarily all) participate in that production (at a cost). This situation incentivizes free riding, since participation involves a cost, but one receives the benefits of the public good nevertheless if others pay that cost in sufficient numbers. If individuals additionally have the capacity to produce, and be sensitive to, a costly signal, quorum signaling systems—that enable individuals to only pay the cost of participation if enough of them are ready to participate—can evolve and be stable under a variety of different conditions. This gives an account of the origins of quorum sensing mechanisms found in populations of organisms as simple as bacteria.

Another example of introducing costs in a signaling game can be found in the work of Santana (2014). The proposed model does not assign costs to messages, but rather to strategies. In particular, the author uses a 4x4x4 Lewis game where if the sender strategy makes use of all 4 possible messages it pays a small cost, whereas if it makes use of only 2 messages, it receives the full payoff. The intuition is that there can be some cognitive or physiological costs to using a larger lexicon (Santana, 2014, pp. 408-409). Using 2 messages necessarily involves pooling, which the author equates to ambiguity. This cost is combined with the presence of contextual information that can potentially be used by the receiver to disambiguate between two states lumped by the sender under the same message. Santana uses agent-based simulations to show that, under a couple of different dynamics (a variant of a birth-death process and a discrete version of the replicator dynamics), the ambiguous use of messages is an optimal strategy that can even take over a population of agents using a precise signaling system.

Agents

David Lewis introduced signaling games as a way to formalize and study coordination problems between two agents: one sender and one receiver. This perspective has been modified and expanded in several ways. First of all, it should be noted that, in simple Lewis signaling games, the agents' existence is only implicit in the modeler's interpretation; what is explicitly represented are two strategies and a utility function. As previously mentioned, other interpretations of these two elements are possible. An important alternative is to think not in terms of individual agents, but in terms of *abstract populations* of agents, where strategies represent an aggregate of the behavior of all individuals in that population and utility represents fitness for survival and reproduction. This forms the base for the evolutionary approach to game theory, popular in theoretical biology (see Maynard Smith, 1982; Nowak and Sigmund, 2004) and introduced to the study of signaling games by Skyrms (1996).

When it comes to how agents are imagined, this marks a move from thinking in terms of two individual agents changing their behavior based on what they consider preferable, to thinking in terms of heterogeneous populations of agents being changed by dynamics external to their will. This perspective also allows for the application of signaling games to the study of communication in organisms simpler than humans, from monkeys, to birds, to even bacteria. When thinking in terms of populations and evolution, one can choose between modeling only one population where each individual plays both roles of sender and receiver, or to have two separate populations, one of senders and one of receivers. While the former is more natural to account for the evolution of communication among agents of the same species, the latter allows for studying forms of inter-species signaling, like those used in cooperative hunting or in prey-predator situations (see Skyrms, 2010, pp. 20-32).

Although the basic setup in a signaling game considers one sender, or one population of senders, and one receiver, or one population of receivers, there are also variations on that theme. Barrett (2006, 2007) has introduced so-called *syntactic games*: signaling models where a receiver handles messages from more than one sender who observe the same state but whose signaling behavior is independent from each other. Barrett observes that, given simple learning mechanisms to update behavior based on obtained payoff in iterative interactions, agents can develop strategies that enable them to communicate perfectly even when the senders can only use a number of messages that is half the number of possible states. Senders do this by effectively partitioning the state space in complementary ways that the receiver can then use to always perform the appropriate action for each state. It is important to stress that this happens despite the fact that senders learn independently from each other.

These games are used by Barrett to argue for hypotheses regarding the evolution of grammar (2006; 2009), conventional kinds (2007), incommensurable languages (2010), basic arithmetic (2012), and epistemic norms (2013). Signaling games with multiple senders and one receiver have also been proposed by Lawry and James (2017) as examples of possible scenarios where sharply delineated partitioning of the state space is less optimal than communication with a certain degree of vagueness. In particular, their results indicate that when a receiver handles messages coming from noisy channels, having multiple vague senders can compensate for transmission errors that would otherwise hamper communication. Skyrms (2009) discusses additional possible setups with multiple receivers, to model situations where one sender coordinates a team of other agents through signaling, chains of agents where some agents play the role of intermediaries between the one who observes the state and the one that carries out the action, and a dialogue setup where both agents signal back to each other. Signaling systems are found to evolve in all of these scenarios.

Communication does not usually evolve between completely isolated agents, neither in homogeneous populations where everyone interacts with everybody else, as abstract population models typically assume. In order to drop these assumptions, one can introduce constraints in the *spatial structure* of agents in a population. Grim et al. (2004) study models with agents fixed on a two-dimensional grid exposed to wandering predators and food sources (think of a simplified coral colony). These agents can be hurt if a predator appears on their position and they are not hiding, or be fed if a food source appears and their mouths are open. They can choose between not signaling or sending one of two signals at each time step, when either nothing is happening, they're being hurt, or they're successfully feeding. They can also receive signals and react to them by doing nothing, hiding, or opening their mouths. Importantly, signals from an agent only reach their direct geographical neighbors.

We thus have a signaling game with three states (nothing happening, being hurt, feeding), three messages (no signal or one of two signals), and three actions (do nothing, hide, open mouth), where agents only interact with a limited number of other agents within the general population. Additionally, agents do not have any direct payoff incentive for cooperation, since they only get punished if themselves are hurt and rewarded if they are fed. The authors find that, either by imitation, recombination, or reinforcement learning, most agents in such a setup develop perfect communication with their neighbors. However, and because this can be achieved by more that one pair of strategies, the whole population does not typically settle on one signaling system, but is split in regional "dialects" with continuously shifting borders. This is an important result, since in simple setups with two agents or abstract populations, the theoretical prediction is that one signaling system always ends up dominating totally. Zollman (2005) achieves similar results in a related, but leaner model of common interest, confirming the importance of considering the spatial structure of a population of agents for understanding the evolution of communication.

Spatial arrangement of agents is especially interesting because of the constraints it imposes on communication. Who exchanges messages with whom is, in the aforementioned models, a function of who is a neighbor of whom. However, most living beings that we know signal, including humans, are typically mobile and not fully constrained in this manner, but they still do not interact with every other agent uniformly; populations typically have heterogeneous *social structures*. Wagner (2009) investigates the impact of different social network topologies on the outcomes of signaling games driven by different adaptive processes. The author replicates and expands on the results of Zollman (2005), and additionally studies agents connected in small-world networks. These are topologies that can be randomly generated to exhibit characteristics found in real-world biological, technological, and more importantly social networks (Watts and Strogatz, 1998).

The results indicate that social networks of the kind investigated promote the development and fixation of signaling systems in a population, even in conditions where this would be less likely to be expected in abstract population models. Small-world networks are found to be "especially hospitable to the emergence of efficient signaling" (2009, pp. 392-393). This kind of work is continued by Mühlenbernd and Franke (2012) who investigate the relation between certain types of agents in terms of where they are located in the network and the role they play in helping either fixate or change a certain signaling convention. More recently, Mühlenbernd (2017) focuses on language change and finds evidence for the so-called weak tie theory, *i.e.* the idea that innovation (new messages, new associations, etc.) is introduced in a system of communication following creative departures from convention by agents in the periphery of the social structure, which then get propagated and made stable by more central members.

Strategies

While states, messages, actions, and utility characterize the game being played, and how one conceptualizes agents introduces additional constraints, strategies characterize how the agents play, should play, or are expected to play the game. As such, they are the element that usually draws the focus of analysis. When studying a signaling game, one is typically interested in which particular strategies are optimal given the setup, and how or if agents end up using them or not. But before one thinks of analysis, it is important to consider the possible kinds of ways strategies can be represented, since different representations enable and motivate different approaches.

Thinking functionally, there are three main types. The simplest are *pure strategies*, where each state is univocally associated with only one message for the sender, and each message is univocally associated with only one action for the receiver. Formally, if T is the state space, M the message space, and A the action space, a pure sender strategy is a function $\sigma : T \to M$ and a pure receiver strategy a function $\rho : M \to A$. These types of strategies can be interpreted as representing either fully deterministic agents or completely homogeneous populations. The advantage is simplicity, especially when it comes to analysis. Unlike with other types, the space of all possible pure strategies, for finite T, M, and A, is also finite, despite the number of possible strategies growing exponentially with the number of states and messages[6]. For example, in a 2x2x2 Lewis game there are only 4 possible pure sender strategies and 4 pure receiver strategies. This makes finding the optimal strategy profiles a tractable problem, as was seen in Section 2.1. Lewis (1969) considers only pure strategies in his original work. They are also the preferred types of strategies in more classic game-theoretical analyses.

A major limitation of pure strategies is that they cannot correctly represent agents that behave non-deterministically (either by choice or by making mistakes), neither can they capture the composition of a heterogeneous population (where different agents in the population may use different strategies). One possible way of modeling these cases is to use *mixed strategies*. These are probability distributions over pure strategies: each possible strategy has an associated probability, a value between 0 and 1 that can be interpreted either as representing the likelihood of an agent behaving in accordance with that strategy, or the proportion of a population that uses that strategy. Formally, $\sigma : \Delta(T \to M)$ and $\rho : \Delta(M \to A)$.[7] Note that, unlike with pure strategies, even for finite T, M, and A, the space of mixed strategies is infinite since probability values are defined in $\mathbb{R}$. Mixed strategies are mostly used in dynamical analyses inspired by biological evolution, with each possible strategy representing a phenotype and natural selection driving changes in the composition of a population via reducing the relative numbers of less successful strategies and increasing the numbers of those with higher fitness. This is the kind of approach introduced into the study of signaling games by Skyrms (1996).

Although mixed strategies are intuitively very suitable for representing heterogeneous populations of agents with each their own pure strategy, it is not clear that the best way to capture the behavior of a stochastic agent is to represent it as switching probabilistically between whole strategies. Another option is to use *behavioral strategies* where probability distributions apply not to pure strategies but only to the choice options for each choice point. Formally, $\sigma : T \to \Delta(M)$ and $\rho : M \to \Delta(A)$, *i.e.* for the sender, each state is associated with a probability distribution over messages, and for the receiver, each message is associated with a

[6] The number of possible strategies is $|M|^{|T|}$ for the sender and $|A|^{|M|}$ for the receiver.

[7] $\Delta(X)$ designates a probability distribution over set X. For a finite X, what we have is, for each $x \in X$, a probability value $P(x)$ such that $\sum_{x \in X} P(x) = 1$.

probability distribution over actions. The space of possible behavioral strategies is still infinite, even for finite T, M, and A, but it is more restricted than for mixed strategies and allows for more compact representations when it comes to implementing computer simulations. Their interpretation as anchored in choice points makes them intuitively suitable for being used in analysis that consider local adaptations of behavior, like processes of imitation or other learning dynamics (see Skyrms, 2010, pp. 83-105).

Analysis

Having set up a signaling game model by putting together all the aforementioned components in a certain configuration, one is left with a theoretical object about which questions can be asked. There are also various ways in which one can go about analyzing and drawing conclusions from a model. Though not strictly being part of the anatomy of a signaling game, certain kinds of models do invite particular methods of analyses in lieu of others, and preferred types of analyses can also condition choices in building models.

Basic aspects of simple signaling games, like those explored by Lewis (1969), can be analyzed using a *static* approach. One possibility is to use a concept that is central in classic game-theoretical analyses of this kind: the *Nash equilibrium*. In general terms, and assuming agents are individuals, two players are said to be in a Nash equilibrium when neither of them can unilaterally change their strategy to achieve a higher payoff. As pointed out by Lewis (1969, pp. 130-141), it is clear that signaling systems are Nash equilibria, but they are not the only ones. This is the case even in binary signaling games: pairs of pooling strategies (as the one illustrated in Figure 2.1c) where the sender does not differentiate between the states (by always sending the same message), and the receiver does not differentiate between messages (by always performing the same action) are in a Nash equilibrium, since neither can do better by unilaterally changing their behavior.

There is, however, an important difference between signaling systems and pooling equilibria: the former are stable where the latter are not. If there is any, even if minute, possibility that one of the players does unilaterally change to a strategy that differentiates, whereas in a signaling system the other player does not have any incentive to switch strategy, in a pooling equilibrium they do. What this means is that, in the simple model we are discussing, players that find themselves in a pooling equilibrium will only remain there in an idealized world where they are perfectly rational, view the other player as perfectly rational, never make mistakes, and there are no perturbations in the environment. Otherwise they will be driven to a signaling system. Note that this depends on the payoff structure and other elements of the game.[8]

The concept of Nash equilibrium is grounded on classic game theory and the view of strategies as representing choices of rational agents. Strategies can, as mentioned before, also be interpreted as descriptions of abstract populations subject to the forces of evolution. In this case, the relevant concept for a static analysis is that of *evolutionarily stable strategy* or *ESS* for short, and was introduced to the study of signaling games by Skyrms (1996). An ESS is a strategy that, if in use by the whole population, cannot be invaded by a small number of agents playing any other possible strategy. Skyrms (1996, p. 90) shows that, for common interest signaling games with the same number of states as signals and actions, in an abstract unstructured population, not only are signaling systems evolutionarily stable, they are also the only ESS's.

Both these concepts help analyze whether a certain combination of strategies is optimal and whether it is stable. They still say nothing about the mechanism by which agents or populations would be driven to play such strategies, and how likely would it be for that to happen. With regards to the former, Lewis (1969) relies on standard assumptions of classic game theory and his own notion of common knowledge: signaling system is stable because agents rationally want to maximize their payoff, thus having no incentive to change strategy, they assume the other agents are rational too and thus do not want to change their strategy, and they believe the other agents ascribe the same beliefs to them. This creates a (potentially infinite) chain of reasoning that is supposed to go on in the minds of agents and requires heavy assumptions on their knowledge of the situation, as well as on their cognitive capacities.

If one relaxes these assumptions, from considering agents with bounded rationality,[9] all the way down to imagining populations of the crudest agents being pushed around by evolutionary forces, a truly *dynamic* study of the system becomes much more attractive. In dynamic approaches, which typically work with mixed or behavioral strategies, one provides a concrete model of the mechanism underlying strategy adaptation and makes calculations of how an initial strategy profile changes under

[8] For example, in a game of pure conflict, pooling equilibria are not only stable, they are the only equilibria of the game (*e.g.* Skyrms, 2010, p. 77).

[9] One of the first to propose analyzing game-theoretic models with agents of limited cognitive capacities in mind was Herbert Simon:

> Broadly stated, the task is to replace the global rationality of economic man with a kind of rational behavior that is compatible with the access to information and the computational capacities that are actually possessed by organisms, including man, in the kinds of environments in which such organisms exist. (1955, p. 99)

said mechanism. This allows studying what kind of attractors there might in the system (the most simple being a fixed point, *i.e.* a strategy profile that does not change under the hypothesized mechanism), and how strategy profiles that do change tend to evolve. Ultimately, one can map out a sample of the space of possible strategy profiles to get an idea of what portion of those profiles ends up in which attractor; this is called calculating its basin of attraction. A dynamic analysis allows for a lot more insight into the details of the adaptive processes that players or populations of agents go through. It also can provide information about the optimal strategies (if there are any), whether these are attainable, and which initial conditions allow for that to happen. It is, therefore, much more comprehensive than a static analysis. The downside is added complexity and a lot of models to choose from.

Skyrms (1996) introduced the use of the *replicator dynamic* (of Taylor and Jonker, 1978) to the study of signaling games. This model attempts to capture, in a very abstract way, the general principles of natural selection: strategies change their relative representation in proportion to their relative success against other strategies in the population. It is thus very suitable to be used in the analysis of signaling game models interpreted in terms of biological evolution. It has also been shown that certain processes of learning by imitation can be represented well by the replicator equation (Björnerstedt and Weibull, 1995; Schlag, 1998), thus making it suitable for interpretation in terms of cultural evolution (Dawkins, 1976; Boyd and Richerson, 1985). The original dynamic comes in two different flavors: a differential equation that represents continuous change through time, and a difference equation that assumes change occurs in discrete generations. While the former is a better representation of how change occurs in nature, the latter facilitates computations for numerical analysis. Both flavors were originally conceived with symmetric games in mind, where the options and payoffs available to each player are the same. This is not the case in signaling games, where sender and receiver have different roles. Two ways of addressing this are possible. One can symmetrize the game by imagining each agent plays sometimes as sender, and sometimes as receiver, thus carrying a combination of strategies, one for each role. Using this approach, one represents the dynamic in one equation for a single population. The alternative is to imagine two populations, one of senders, and one of receivers, co-evolving with each other, which results in a representation of the dynamic in terms of a system of two equations.

The dynamic analysis of binary signaling games is consistent with the static insights: signaling systems are the only stable attractors in both the one-population and two-population replicator dynamics (Skyrms, 1996; Huttegger, 2007a; Hofbauer and Huttegger, 2008). Complications arise, however, as we soon as we deviate from this simple game in various ways (see Huttegger, 2007a; Skyrms, 2010, pp. 63-72). Interestingly, some of these variations on the simple signaling game that make the evolution of signaling systems be less then a guaranteed certainty perform differently under a variant of the replicator equation. The *replicator-mutator dynamic* is an extension that incorporates an aspect of evolution that is so far missing: mutation. Signaling systems can be shown to be even more robust under this dynamic than under the simple replicator (Hofbauer and Huttegger, 2008). Other variants are possible. For example, Franke and Correia (2018) derive a discrete form of the replicator dynamic for behavioral strategies where agents learn by imitation but are subject to imprecision, both when observing the behavior of others as well as when trying to mimic it. In certain sim-max games, under this *imprecise imitation dynamic*, strategies bearing characteristic marks of vagueness are observed to evolve and be stable.

There are also dynamic models that are based on higher assumptions of rationality. Gilboa and Matsui (1991) propose a *best response dynamic* where agents have high awareness of the game and always select the optimal choice of behavior given the game and state of the population. Franke, Jäger, and van Rooij (2011) study a variant of this model where limits to the agents rationality can be introduced, and suggest that bounded rationality can, in a sim-max game, induce equilibria where agent's strategies have characteristics of vague signal use. A more detailed overview of evolutionary dynamics is provided by Huttegger, Skyrms, Smead, et al. (2009).

Population dynamics describe adaptive processes at a very abstract level. Although some of the approaches discussed can capture general principles of learning, it can be interesting to study more specific mechanisms and their implications for the development of successful communication. One of the most well studied learning mechanisms is *Roth-Erev reinforcement* (Roth and Erev, 1995). The main idea it tries to capture is that an agent makes a choice with a probability proportional to the accumulated rewards attained from making it in the past. The usual illustration of the process portrays each agent as having an urn per choice point (states for senders, messages for receivers) that contains balls colored by possible choice (one color per message for the sender, one per action for the receiver). Actual choices are made per choice point by randomly drawing a ball from its urn and selecting the choice associated with the color of the drawn ball. Additional balls are then added back proportionally to the payoff received from making that choice. This reinforces successful choices by increasing the probability of making them again in the future, and concomitantly reduces the probability of choosing any of the alternatives.

This type of learning mechanism is highly conducive to the development of successful signaling in simple Lewis signaling games (Barrett, 2006; Argiento et al., 2009; Catteeuw and Manderick, 2014) and has been generalized and extended in several ways. Barrett and Zollman (2009) add the ability for agents to forget past experiences and find that this has a positive effect on learning. Alexander, Skyrms, and Zabell (2012) add the possibility of new messages being invented during the learning process, which not only allows for modeling something that happens in natural language, but is also observed to have benefits for the development of optimal communication. O'Connor (2014b) proposes a *generalized reinforcement learning* rule for sim-max games where successful choices reinforce not only the original choice points but also similar ones; this is found to speed up learning and can be used to explain the evolution of vagueness. Franke (2014a) combines some of these variants to propose a mechanism for the development of compositional signaling. Catteeuw and Manderick (2014) compare Roth-Erev reinforcement to Q-learning and automata to find that all three mechanisms efficiently learn signaling conventions. Yet another family of models of individual learning that has been successfully tested in signaling games is *artificial neural networks* (*e.g.* Davis, 2017). There is a large number of general learning models and variations proposed in the literature. I give here emphasis to those developed specifically for signaling games, but others, coming from areas like behavioral economics and experimental psychology, might bring additional relevant insights (see Skyrms, 2010, pp. 83-105).

Raising the level of rationality of the agents even further, one can consider and attempt to model the use of strategic reasoning. Note that, although Lewis (1969, pp. 24-32) explores the use of complex reasoning with higher-order expectation as ways to solve coordination problems, he does not actually formalize them nor does he study how they play out in a concrete model. This has been done more recently in work that aims to model mechanisms of pragmatic reasoning in language. An overview of the approach and the related literature is given by Franke (2017).

These learning mechanisms are typically modeled for the case of two individual agents interacting repeatedly with each other, and implementations are almost as abstract as population dynamics. Learning occurs at the agent level, but one might also be interested in studying the case where an agent learns from various individuals in a heterogeneous population. One can then choose to model each agent individually, maintaining and updating independent strategies and internal states. These *agent-based models* are typically implemented and run as computer simulations of actual play. Examples of these include the aforementioned research of Grim et al. (2004), Zollman (2005), Wagner (2009), Mühlenbernd and Franke (2012), and Mühlenbernd (2017).

An additional example can be found in the work of Franke, Jäger, and van Rooij (2011), who explore a variant of learning by fictitious play (G. W. Brown, 1951) in which agents have a limited memory of past interactions. The outcome of a large number of agents playing a sim-max game in repeated interaction with each other show, after a period of convergence, a stable and persistent variation in each agent's evolved strategy that, the authors argue, induces characteristic marks of vagueness at the aggregate population level. Agent-based models are popular in social science and gaining adoption in many other areas (see Wilensky and Rand, 2015). They are used to study complex phenomena by establishing a connection between low-level mechanisms of interaction and the high-level emergent patterns they can induce in certain environments. As such, they provide a method of analysis that brings additional insight into the interplay of all the factors that influence the development of successful signaling.

Given a choice of static or dynamic approaches, one can still further opt between two ways of drawing out conclusions. If the game model and the dynamics (in case that is the chosen approach) are specified in mathematical equations, it is sometimes possible to determine equilibria, basins of attraction, or other properties of the system using *symbolic* analysis. This involves calculating exact solutions to mathematical formulations of the investigated properties in relation to the equations that define the system. A good example of a fully symbolic analysis can be found in the aforementioned work of Argiento et al. (2009). Although this type of approach has the advantage of being precise, it requires a full mathematical specification of the model. Therefore, it can easily become either too difficult, unfeasible, or even impossible, as models are complicated beyond the most basic.

In those situations one can resort to *numerical* analysis and calculate statistical approximations to the properties one is interested in investigating. This is usually done with the aid of computer simulations. An example is a standard method for estimating the attractors of a dynamic by running a kind of Monte Carlo experiment: taking as domain the space of possible population proportions of all possible strategy profiles, one can randomly sample a large number of values from that domain, use the model of the dynamic to calculate the changes in each population proportion until they stabilize, and determine how many converged into which strategy profile. This was originally used by Skyrms (1996), who found that all initial population proportions in a simulation of the simplest signaling game converged, under the replicator dynamic, into one of the two possible signaling systems, going to each in approximately equal proportion. This does not prove that no other attractors exist, but it can raise our confidence that that is the case.

Although numerical methods are not able to prove universal claims about the models, they can provide existence proofs. For example, Huttegger, Skyrms, Smead, et al. (2009) use numerical methods to show that the replicator dynamic can lead some population proportions in a simple signaling game with three states, messages, and actions, into partial pooling equilibria. The choice between symbolic and numerical methods is conditioned by how the model is formalized, the researcher's ability or inclination towards one or the other approach, available resources (Monte Carlo simulations were not so easy to perform before the advent of powerful personal computers), but also to a certain extent tradition: research on signaling games in economics, stemming from Spence (1978), tends to use the symbolic approach; in philosophy, work stemming from Skyrms (1996) tends to prefer numerical analysis; in biology, where researchers focus primarily on costly signaling (Grafen, 1990; Maynard Smith, 1991), there is more of a balance with a slight skew towards numerical methods.

Unlike symbolic analysis, numerical methods, especially when applied in dynamic analysis, typically generate a large number of data points that subsequently need to be aggregated and summarized to yield meaningful results. In Monte Carlo and other numerical investigations of the dynamics of a signaling game model, this involves collecting statistics either (or both) at the end of the simulation and in the intermediate steps. This can range from simple values, such as the proportion of simulations that converged to each strategy profile, to more complex *metrics*. One example of a metric is the *quantity of information*. Skyrms (2010, pp. 33-47) defines the quantity of information carried by a signal about a state as the logarithm of the ratio between the probability of the state conditional on getting the signal and the prior probability of that state. If sending the signal does not alter the probability of a state obtaining, then the signal carries not information, otherwise the metric quantifies how much information is carried by the signal. This can be aggregated into an overall quantity of information of a signal by averaging over all states. A similar metric can be defined with actions, rather than states, in mind.

Wagner (2012) makes use of this metric to show that, even in a signaling game model that never converges to equilibrium, information is still partially transferred as sender and receiver strategies continuously change and adapt to each other. Godfrey-Smith and Martínez (2013) and Martínez and Godfrey-Smith (2016) propose information transfer should be measured not for messages but for strategy profiles as a whole in terms of the amount of information between states and actions. Using this metric, they corroborate and expand upon Wagner's results. Other metrics are possible. For example, Franke and Correia (2018) define entropy, convexity, and additionally track expected utility to gain insight into certain properties of strategies and how they change while the system evolves under their imprecise imitation dynamic. As with other aspects of signaling games, which particular metrics are relevant depends on the game under study and the questions one is trying to answer.

Part II

Later Wittgenstein and signaling games

Chapter 3

Room for systematicity[*]

Does Wittgenstein's metaphilosophy invite us to abandon systematicity in philosophy? What does his practice further reveal? In this chapter, I look into Wittgenstein's remarks on method and their consequences for the legitimacy of making use of the signaling games framework.

Before reflecting on whether or not Wittgenstein's metaphilosophy leaves room for systematicity, it is important to clarify what is meant by the term. A strong notion of systematicity is related to the idea of system-building. To be systematic in philosophy, in this sense, is to work towards devising a coherent picture of how everything hangs together, spanning and connecting all areas of what can be thought. A systematic philosopher of this kind attempts to establish a firm foundation, a small number of basic principles from which to derive a unified understanding of, or approach to, a whole host of philosophical problems in areas from ontology, to epistemology, language, mind, aesthetics, ethics, politics, and more. We can recognize this type of ambition in authors like Plato, Aristotle, Descartes, Leibniz, Spinoza, Hume, Berkeley, and Kant, to name a few examples. This is not the kind of systematicity relevant for the thesis defended here. My interest lies in the connections between Wittgenstein's later picture of language and the framework of signaling games. Given that signaling games do not aim to be a philosophical system in this broad sense, the question of how the approach is in line with Wittgenstein's thought does not hinge on this kind of systematicity.

Signaling games are, however, systematic in a different sense. Setting up a signaling game model involves formally defining a number of elements and specifying how they are interrelated. The framework is not a philosophical system in that it does not aim to be a description of reality, let alone to explain how everything

[*] Some of the content in this chapter (mostly Section 3.1 and Section 3.2) overlaps (with minor modifications) with the publication entitled "Analysis and Explanation in the *Philosophical Investigations*" (Correia, 2019).

hangs together. However, it fits a weaker notion of systematicity in that it obeys a certain method: some core elements are always specified (*e.g.* states, messages, actions), solution concepts fit within some well-defined principles (*e.g.* utility maximization), and results follow from assumptions according to mathematical formulas and calculations. If one wants to see how this approach fits with Wittgenstein's later philosophy, it is important to reflect on whether this kind of systematicity would be in line with his ideas regarding methodology.

Each of the camps outlined in Section 1.1 has a different understanding of which methods Wittgenstein supposedly rejected or avowed. In this chapter, I defend that Wittgenstein's issues with methodology are not for or against particular methods, but are instead related to the attitudes we have towards them, *i.e.* to what we believe we are doing and can achieve with them. I argue that Wittgenstein's pragmatist attitude tolerates a plurality of methods, and a variety of them are in fact used in the *Philosophical I*nvestigations. Some arguments in the book involve setting up thought experiments with artificial scenarios of language use. I argue that signaling games can be interpreted as similar to these toy language-games in many respects, and thus can be seen as methodologically in line with his practice. I make these arguments by first reflecting on what Wittgenstein had to say about certain aspects of philosophical practice that are related to its quintessential philosophical method of analysis (Section 3.1). Next, in Section 3.2, I look into the notion of semantic explanation and how it can inform a different attitude towards particular methods in philosophy. In Section 3.3 I explore one method often used in the *Philosophical Investigations*: the use of toy language-games as thought experiments. Finally, in Section 3.4, I explore in detail the connection between toy language-games and signaling games.

3.1 In pursuit of chimeras

The core of Wittgenstein's observations on philosophy and method in the *Philosophical Investigations* can be located in §§89-133. An important part of those remarks engages in criticism of a certain way of doing philosophy. This is accompanied by what one could call a diagnosis of the motivations underlying the tendency to follow that path. Wittgenstein sets off this discussion by asking the question "In what way is logic something sublime?" (§89a). The observations that follow are, however, not specifically about logic as a formal tool or field of research. They are more concerned with logical thinking as a method, and more generally about what can drive us into believing that this kind of thinking has a "peculiar depth":

> Logic lay, it seemed, at the foundation of all the sciences. – For logical investigation explores the essence of all things. It seeks to see to the foundation of things, and shouldn't concern itself whether things actually happen in this or that way. (2009, §89b, pp. 46e-47e)

That this conception is to be questioned is already given away by the qualification "it seemed". Logic serves here as a representative of an attitude common in philosophy: the quest for essences, foundations, underlying structures. It is also evocative of the method of philosophical analysis.

The paragraph is followed by an example from the writings of Augustine about a typical philosophical question ("What, then, is time?") and subsequently by remarks that draw attention to the kind of urges triggered by this type of questions:

> We feel as if we had to *see right into* phenomena: yet our investigation is directed not towards *phenomena*, but rather, as one might say, towards the '*possibilities*' of phenomena. What that means is that we call to mind the *kinds of statement* that we make about phenomena. So too, Augustine calls to mind the different statements that are made about the duration of events, about their being past, present or future. (These are, of course, not *philosophical* statements about time, the past, the present and the future.)
> Our inquiry is therefore a grammatical one. (2009, §90, p. 47e)

These remarks go against the idea that philosophy investigates not surface language, but something beyond it. Wittgenstein is saying that, when exploring typically philosophical questions, even though one can believe to be making considerations directly about phenomena, such as time, one ends up actually reflecting on statements that are made about the phenomena. This is partly what is meant when it is said that the inquiry is a grammatical one.[2]

Augustine continues his analysis as follows:

> Yet we speak of a "long time" and a "short time," though only when we mean the past or the future. For example, we say that a hundred years is a long time ago or a long time ahead. A short time ago or a short time ahead we might put at ten days. But how can anything which does not exist be either long or short? For the past is no more and the future is

[2] It is important to note here that Wittgenstein's use of 'grammar' and its cognates relates to meaning, rather than to syntax or other aspects one would typically associate with the term given its use in contemporary linguistics. For more on Wittgenstein's use of these terms, and his idea of grammatical investigations, see McGinn (2011).

> not yet. Surely, then, instead of saying “It is a long time” we ought to say of the past “It was a long time” and of the future “It will be a long time.”
>
> My Lord, my Light, does not your truth make us look foolish in this case too? For if we speak of a long time in the past, do we mean that it was long when it was already past or before it became the past and was still the present? It could only be long when it was there to be long: once it was past it no longer was, and if it no longer was, it could not be long. (Augustine, 1993, Book XI, §15)

We see here how Augustine goes back and forth from considerations of a linguistic nature regarding what we *say about* time and when we *call* time long or short, to ontological considerations of how time *can be* long or short. Underlying this approach is an implicit understanding of the bearing these statements have on the characterization of the phenomena. The relevance of exploring the former in order to understand the latter implicitly buys into a picture of language as strongly connected with other kinds of entities, namely the ones that the philosopher is interested in knowing more about. It seems to presuppose at least two things: first, that because we have a word ‘time’ there is a Something (§261, §293) to which the word corresponds; second, that the statements we make about ‘time’ must to some extent capture the properties of that Something. Driven by this picture, philosophers can be lead to rashly hypostatize linguistic entities, to think that things like language, propositions, thought, have a hidden essence that “an analysis is supposed to unearth” (2009, §92, p. 48e).

This implicit understanding of language is strongly tied to an urge that goes hand in hand with this kind of philosophical method. The statements we make about phenomena seem to lack exactness. We ask ourselves, like Augustine: ‘What exactly do we mean when we say time is short?’ Our craving for precision, coupled with the image of meaning as a Something, can lead us into thinking that these statements can be further analyzed and expressions completely clarified. Wittgenstein’s allusion to this problem (§§91-92) is followed by examples of what could be said about propositions (§§93-94) and thought (§95), and how these statements can be muddled by the “tendency to assume a pure intermediary” (2009, §94, 9. 48e). This idea can then lead us back into believing that what we really need to grasp is the essence of language, *i.e.* “the order existing between the concepts of proposition, word, inference, truth, experience, and so forth” (2009, §97, p. 49e). We puzzle about vagueness in our everyday language (§§98-100) and then project the requirements of exactness and generality back into reality (§§101-107, §§110-115). Wittgenstein seems to be suggesting that this craving for exactness and the picture of language as anchored in a form of correspondence reinforce each other in sending philosophers in "pursuit of chimeras" (2009, §94, p. 48e), and lead them astray into philosophical confusion:

Remark 3.1.1 *A picture of language anchored in correspondence, together with a craving for exactness, lead us into philosophical confusion.*

The relevance of this diagnosis for philosophical practice can be better understood if we reflect on a particular method. Analysis can be seen as one of the most quintessential in philosophy. Michael Beaney (2016) argues that, although metaphilosophical conceptions and definitions of the method have varied through the ages, there are aspects of philosophical practice that can be subsumed under a broad conception of analysis, and these can be characterized by three main perspectives. One is the *regressive* conception of analysis, which conceives the method as aiming at "working back to first principles by means of which something could then be demonstrated" (2016). Another perspective is to think of analysis as *decompositional*, *i.e.* as breaking down concepts (or propositions, or linguistic complexes, or facts) into their simpler constituents. Yet another conception emphasizes a *transformative* (also called interpretative) dimension, in that performing analysis involves a type of translation from one form into another. These are not to be conceived as distinct characteristics of analysis, but rather as intertwined aspects of the method. To what degree each of these aspects or perspectives of analysis comes to the fore is something that varies with different philosophers, both in their practice, as in their explicit conceptions of the method.

In order to illustrate these three aspects, let us take as an example the traditional analysis of knowledge as justified true belief[3]. We can say that this analysis is regressive in the sense that it motivates one to orient one's considerations about knowledge recursively as considerations about justification, truth, and belief, *e.g.* if you want to inquire whether someone knows X you should ask whether she believes X, is justified in believing it, and whether or not X is true. The analysis is also decompositional, in the sense that it purports to expose the internal structure of the concept of knowledge in terms of these other concepts, which are thus seen as constituting it. The transformative aspect relates to the idea that 'justified true belief' is like a translation of 'knowledge', an interpretation of it in a different form

[3] I make absolutely no claims regarding whether or not the analysis is good, I merely introduce it as an example of an analysis that hopefully is familiar enough to motivate a better understanding of the aforementioned aspects of analysis.

with the same characteristics, which should thus be semantically interchangeable with it.

Although no explicit critique of analysis in this sense exists in the *Philosophical Investigations*, there are passages that raise issues with each of the aforementioned three aspects. Since they are intertwined with passages that I see as characterizing the proposed alternative, I will delve into these in more detail in the next section. Before that, I would like to focus on another point. The issues with each of the three aspects of analysis in a broad sense stem from what Wittgenstein sees as a misguided picture of language, and with an attitude towards philosophical methodology that goes hand in hand with that picture. I maintain that Wittgenstein's problem is not with a particular method like philosophical analysis, but rather with the attitude of idealizing the power and purpose of any method. A passage quoted earlier continues as follows:

> Our inquiry is therefore a grammatical one. And this inquiry sheds light on our problem by clearing misunderstandings away. Misunderstandings concerning the use of words, brought about, among other things, by certain analogies between the forms of expression in different regions of our language. – Some of them can be removed by substituting one form of expression for another; this may be called 'analysing' our forms of expression, for sometimes this procedure resembles taking a thing apart. (2009, §90b, p. 47e)

This passage seems to strike a conciliatory tone. Wittgenstein is suggesting that one *can* clear misunderstandings away by doing something which "may be called 'analysing'".

Wherein lies the problem, then? The section that follows §90 immediately reveals that troubles start to arise when we expect too much of the method, when we believe that, because *once* an analysis helped clear up *one* misunderstanding, we can continue analyzing until we clear up *all* possible misunderstandings:

> But now it may come to look as if there were something like a final analysis of our linguistic expressions, and so a single completely analysed form of every expression. That is, as if our usual forms of expression were, essentially, still unanalysed; as if there were something hidden in them that had to be brought to light. As if, when this is done, the expression is completely clarified and our task accomplished.
>
> It may also be put like this: we eliminate misunderstandings by making our expressions more exact; but now it may look as if we were aiming at

> a particular state, a state of complete exactness, and as if this were the real goal of our investigation. (2009, §91, pp. 46e-47e)

Wittgenstein takes issue not with analysis in particular, but with overvaluing its merits or overestimating its potential. It is acceptable to think of analysis of an expression as regressive, as long as we do not expect to reach a final completely analyzed form of that expression; to see it as decomposing an expression, as if taking a thing apart, as long as we do not see this as recovering the real meaning of the expression; to substitute one expression for another in a particular situation, as long as we do not get trapped in the illusion that the latter form is therefore always better than the former and that one can or should substitute it in every situation. The rejection is of ideals of finality, essentialism, and exactness. I believe that Wittgenstein would reject these ideals for any philosophical method, not just analysis. We should thus keep the following remark in mind when reflecting on systematicity in philosophy.

Remark 3.1.2 *Philosophy should stay away from the ideal that linguistic expressions can have a hidden single final exact analyzed form.*

The rejection of this ideal is closely related to Wittgenstein's separation between philosophical and scientific questions, and his rejection of theory-building and explanation. The most notable passages on this issue can be found in §109:

> It was correct that our considerations must not be scientific ones. [...] And we may not advance any kind of theory. There must not be anything hypothetical in our considerations. All *explanation* must disappear, and description alone must take its place. (2009, §109, p. 52e)

It has often been noted (*e.g.* Gruender, 1962; Baker and Hacker, 1980; Ben-Menahem, 1998; Glock, 2007; Hacker, 2012) that the notions of theory, hypothesis, and explanation at stake in these remarks relate to scientific method and practice, especially in the natural sciences. Wittgenstein thought that philosophical problems are of a different nature than scientific problems, and thus require a different approach. It is important to dissect the characteristics that motivate the rejection of scientific method as suitable for addressing philosophical problems. Explicit statements on this are unfortunately lacking in the *Philosophical Investigations*. However, we can find more enlightening remarks in the *Blue Book*, for example the following:

> Our craving for generality has another main source: our preoccupation with the method of science. I mean the method of reducing the explanation of natural phenomena to the smallest possible number of primitive

> natural laws; and, in mathematics, of unifying the treatment of different topics by using a generalization. Philosophers constantly see the method of science before their eyes, and are irresistably tempted to ask and answer questions in the way science does. This tendency is the real source of metaphysics, and leads the philosopher into complete darkness. I want to say here that it can never be our job to reduce anything to anything, or to explain anything. Philosophy really is 'purely descriptive'. (Wittgenstein, 2002, p. 18)

It is clear from this passage that, when saying that philosophy should not aim to explain anything, Wittgenstein is identifying explanation with this portrayal of the method of science as reducing explanations of phenomena to a small number of laws and striving for generalizations.

Note that the rejection of the scientific approach, has to do with the nature of the problems, rather than with an arbitrary classification of something as philosophy or science based on conventional or societal criteria. Thus, if a physicist decides to conduct an inquiry into "What, then, is time?", he is likely under the same predicament as the philosopher investigating the same question. As mentioned before, characteristically philosophical questions are about meaning, and problems arise from misunderstanding language and being driven to hypostatize words. A further passage is evidence of how Wittgenstein sees the relation between this issue and the attitude of philosophers towards the method of analysis:

> Philosophers very often talk about investigating, analysing, the meaning of words. But let's not forget that a word hasn't got a meaning given to it, as it were, by a power independent of us, so that there could be a kind of scientific investigation into what the word *really* means. (Wittgenstein, 2002, pp. 27-8)

Meaning is, for Wittgenstein, not something objective and fixed, and thus cannot be investigated like the objects of study of scientific inquiry.

Another way to understand the dichotomy is by looking into the opposition between explanation and description that is present in the *Philosophical Investigations*. Ben-Menahem (1998) points out a number of ways in which these concepts are different. They can be condensed, I believe, into three main points. First, scientific explanations purport to be objective. Natural laws are supposed to be independent of the scientist's idiosyncratic personal history and individual experience of the phenomena. Second, scientific theories are built on hypotheses that involve theoretical entities and nomological relations. Third, such theories typically result from inductive generalization and are presented as universal and atemporal. Wittgenstein's view of philosophical problems (as pointed in Remark 3.1.1, p. 57) is that their nature is linguistic, they arise in the context of particular language-games and are characterized as confusions or misunderstandings. They are, therefore, situated and their resolution requires a first-person perspective, not an objective one. Regarding the second difference, evoking theoretical entities and pursuing nomological relations is undesirable in philosophy since it amounts to falling prey to the kind of urges that lead to philosophical problems in the first place (see again Remark 3.1.1, p. 57). With regards to the search for universal atemporal truths, this again conflicts with the nature of philosophical problems as afflicting individuals which occupy a particular position in space and time. Some of these positions will become clearer when we discuss an alternative conception of explanation in Section 3.2. For now, what is important to retain is Wittgenstein's negative position which can be summarized in the following remark:

Remark 3.1.3 *Philosophy should not aim for objectivity, postulate hidden entities, or look for universal atemporal truths.*

If this interpretation is right, Wittgenstein takes issue with a number of aspects of philosophical methodology that are not specific to analysis. First, problems start with a picture of meaning as a Something that has properties defined independently of us. Second, philosophical inquiry can fall prey to the urge of exactness and the quest for objectivity, lead astray in pursuit of the chimeras of universal atemporal truths. Wittgenstein notoriously eschewed explicitly naming the targets of his criticism. In the rare occasions where he does, some ideas of Frege, Russell, and his younger self are mentioned in a negative light. This, together with the ideas criticized, the historical context, and Wittgenstein's own path and influences in philosophy, should make it clear that a good example of the attitude under scrutiny can be found in the project of early analytic philosophy (Baker and Hacker, 1980, pp. 259-293). But the criticism runs deeper, and some of the aspects just mentioned can be found in other schools of thought. Wittgenstein's metaphilosophical remarks target not a particular method, like philosophical analysis, but rather a broader attitude towards philosophy that has an influence on how any method is used. Analytic philosophy has changed since its early days, but this attitude can still be identified in contemporary authors both within and beyond it.[4]

[4] It would be inappropriate to point fingers without proper argumentation, and to delve into that would divert us greatly from the point at hand. These arguments are therefore relevant as a criticism against analytic philosophy today only insofar as the reader agrees with this diagnosis.

What does Wittgenstein advise philosophers alternatively do? Along with these remarks come suggestions for different tasks for philosophy. The confusions that arise from misunderstanding language and craving for exactness are undesirable because they "send us in pursuit of chimeras" and "prevent us in all sorts of ways from seeing that nothing extraordinary is involved" (2009, §94, p. 48e). If philosophy should not build theories, theses, or hypotheses, is there anything left for philosophers to do? A number of observations in §§89-133 seem to suggest that philosophy should be confined to clearing misunderstandings away (§90), describing rather than explaining (§109, §124, §126), bringing words back to their everyday use (§116), and throwing light on features of our language (§130) aiming to make philosophical problems completely disappear (§133). These remarks point to the idea that philosophy's task should be to dismantle the problems that arise out of the misguided picture of language and associated urges that lead philosophers into confusion.

Does Wittgenstein propose a concrete method for such a task? As discussed in Section 1.1, opinions vary. Many interpreters of Wittgenstein focus on the *Philosophical Investigations* as defending a particular philosophical method or having an overarching strategy. Defenders of the so-called orthodox interpretation (started by Baker and Hacker, 1980) describe it as grammatical clarification, which has been characterized as "marshalling an ordered array of familiar rules (grammatical rules) for the use of words" (Hacker, 2012, p. 4). This can be supported by passages like §122 and §127. This interpretation is close to the idea of dissolving philosophical problems by studying and describing ordinary uses of language in more detail, which is characteristic of the so-called ordinary language philosophy (*e.g.* Ryle, 1962; Austin, 1962). Motivation for this approach can come, for example, from §116. Others (*e.g.* Wisdom, 1953; Baker, 2004) have made a parallel with psychoanalysis, insisting that Wittgenstein viewed philosophy as a personal therapeutic activity. A remark that can inspire such a view is §255. A related interpretation (*e.g.* Fogelin, 1976; Stern, 2004; Plant, 2004) sees in the *Philosophical Investigations* a defense and example of the methods of Pyrrhonian skepticism. Part of §133 certainly raises such motifs. And the list goes on (see Glock, 2007).

One could argue about the particular nuanced differences between each of these proposed methods or strategies, and which of them is *the* one that Wittgenstein truly espoused. But the simplest explanation for the variety of interpretations regarding Wittgenstein's methodology, and for the ease with which each author can find passages supporting their own view, is that elements of *all* of these views might be present in the *Philosophical Investigations*. And this is not the result of accident or sloppiness on the author's part. Wittgenstein explicitly avows for methodological pluralism in philosophy:

> There is not a single philosophical method, though there are indeed methods, different therapies, as it were. (2009, §133d, p. 57e)

To look for a unifying method or overarching strategy is to ignore not only this passage, but also the variety of elements of different methods and argumentative strategies that can be found in the book. It is clear that the metaphilosophical remarks put a strong emphasis on defending that philosophy should be confined to clearing linguistic misunderstandings away.

This recommendation, what Wittgenstein explicitly defended, does fit better with the therapeutic, Pyrrhonian, or quietist interpretations. But is practice reveals something more. Even these mostly negative views are anchored in a particular picture of language and meaning, which itself is not independently supported. These constitute what Plant (2004) would call Wittgenstein's minimal dogmatism. There could be a disconnect between what Wittgenstein defends philosophers should do and his own philosophical practice that he himself did not realize. This possibility leaves a bit more room for drawing some positive views from the *Philosophical Investigations*. Making the parallel between explanations of meaning and philosophical methods, as I argue in the next section, can help us see this even more clearly.

3.2 Explanations of meaning and philosophical method

Wittgenstein's paradigmatic examples of misguided philosophical pursuits are driven by questions like 'What is language?' or 'What is a proposition?' (see §92). As discussed in the previous section, the inquiry these types of questions lead to can be fruitless if one uncritically relies on a certain picture of language. In particular, Wittgenstein criticizes the idea that words like 'language' and 'proposition' already correspond to a Something, and it is the nature of this Something that is the subject of investigation. If one sees meaning as this kind of correspondence, those questions are no different from the questions 'What is the meaning of the word 'language'?' and 'What is the meaning of the word 'proposition'?', for the meaning of these expressions is the Something one is interested in investigating..

If the assumption of correspondence is left unchecked, one will tend to go down the same path of inquiry when considering either the first form of the questions or the second one. But, if one sees meaning as related to use (see Section 4.2), it is possible to approach them differently:

> "The meaning of a word is what an explanation of its meaning explains." That is, if you want to understand the use of the word "meaning", look for what one calls "an explanation of meaning". (2009, §560, p. 158e)

Similarly, if you want to understand the meaning of the word 'language' or 'proposition', look for what one calls an explanation of these words.

In general, then, Wittgenstein's recommendation is to replace questions of the form 'What is X?' with questions of the form 'How do we usually explain X?'. This kind of methodological advice is what connects Wittgenstein with the so-called *ordinary language philosophy* inaugurated by authors like Gilbert Ryle and J. L. Austin. I am not interested in defending whether or not this is an adequate interpretation here. What I am interested in is the link between explanations of meaning and philosophical method. The ways that we explain the use of words to one another are, according to the view just adumbrated, relevant to how philosophers should go about understanding those words. This is, furthermore, suggested to be a better way of reflecting on the philosophical problems connected with those words and to avoid falling prey to the typical misunderstandings that Wittgenstein warns of.

Before we get started, it is important to clarify the notion of explanation of meaning. A clear characterization is given by Baker and Hacker (1980, ch. 2). Wittgenstein talks about two ways in which meaning can be clarified between two agents: training and explaining. Training is the most basic way of teaching the use of words. In §5, it is said that it is by training, not explaining, that a child learns to talk. Part of this training can involve "the teacher's pointing to the objects, directing the child's attention to them, and at the same time uttering a word" (2009, §6, p. 7e). This is what Wittgenstein calls "ostensive teaching of words". Other examples of activities learned by training include using a chart (§86), making calculations according to an algebraic formula (§189), following a signpost (§198), and obeying an order (§206), which are all instances of rule-following. Although a more explicit definition is lacking, one can surmise that training amounts to learning how to perform an action in response to certain linguistic expressions or symbols.

Explanations, on the other hand, are ways of teaching the use of words by means of other words. Examples include ostensive explanation[5] (*e.g.* §§28-36, §73), giving examples (*e.g.* §68, §71), referring to samples (*e.g.* §50), or sentence paraphrasing (*e.g.* §20, §§60-64). For the sake of brevity, and alignment with the source text,

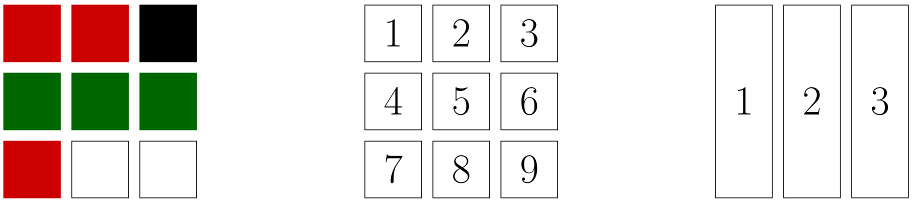
[5] Also called ostensive definition.

in this section I will talk about explanations simpliciter, but it should be noted that the remarks in here are about explanations of meaning, rather than nomological explanations (of the scientific kind). This is following the distinction made in the previous section. Although Wittgenstein rejects nomological explanations as appropriate for philosophy, explanations of meaning are different and, I argue, relevant to fully understanding Wittgenstein's attitude towards philosophical methods.

Going back to Beaney's three aspects of analysis, I would like to first draw attention to §§19-32, which discuss a number of questions closely linked with the transformative aspect of analysis. The initial sections question the idea that one can reveal the meaning of the words used in the builders' language of §2 by translating them into a different form. Should we say that, in that language-game, the call 'Slab!' actually means 'Bring me a slab!'? That is one way we could explain the use of the call to someone who was not familiar with that particular language-game. But such a translation would only help someone who already knows how to use the other words in the translated form like 'bring' or 'me', *i.e.* someone who has played other language-games. To the hypothetical primitive builder that only plays the language-game of §2, the translated form would actually be incomprehensible because he does not know how to use those other words.

The point of these remarks is two-fold. First, to remind us that just because we can translate a linguistic expression into another that does not make the latter the *real meaning* of the former. There is nothing more fundamental to the builder than the call 'Slab!', and there is no way in which he means 'Bring me a slab!' when he uses it. This goes clearly against transformative ideals of unearthing true or fundamental forms of linguistic expressions. One can think of a large part of Frege's philosophy (explicitly mentioned in §22) or Russell's theory of definite descriptions (1905) as examples of projects that implicitly or explicitly held to that ideal. The second point of these remarks is to draw attention to how verbal explanations of meaning are always anchored in particular language-games. This is very explicitly stated for ostensive definitions in §§30-31, where the motto is that "an ostensive definition explains the use – the meaning – of a word if the role the word is supposed to play in the language is already clear." (2009, §30a, p. 18e) And to know the supposed role of a word is to know how to play a certain language-game.

Wittgenstein uses, in §31, the example of chess to drive home the point. The short explanation 'This is the king' (while holding or pointing to a chess piece) only helps the other if he already knows how to play, but does not know which piece is supposed to be the king in that particular board. For a less informed partner, the explanation 'This is the king; it can move in this-and-this way' can help if the other



(a) Arrangement. (b) Interpretation schema 1. (c) Interpretation schema 2.

Figure 3.1: Arrangement of colored squares from §48 and interpretation schemas.

has played other games and knows what a piece is, what a move is, how these are coordinated (*e.g.* taking turns), and so forth. He needs to already know how to play games and how to use the words given in the explanation. Ultimately, this point should follow from the simple fact that, by definition, explanations of meaning always involve the use of words, and words are always learned in the context of particular language-games.

This point has implications for the decompositional aspect of analysis as well. The notions of simple and composite, the cornerstones of the idea that a concept can be broken down into its constituent parts, are put into question in §§47-48.[6] In the current context, I want to draw attention to Wittgenstein's remarks that these two notions are always relative to each other: what is simple depends on what kind of compositeness one is interested in, and this can vary depending on the language-game one is playing. This point comes back more markedly in §§60-64, where Wittgenstein draws again on the language-game introduced in §48. Consider Figure 3.1. Using the order represented by the numbers in interpretation schema 1 (Figure 3.1b), and the letters 'R', 'G', 'W', and 'B' for, respectively, the colors red, green, white, and black, one could describe the arrangement in Figure 3.1a by the sentence 'RRBGGGRWW'. In this language-game, it might seem obvious that each colored square is a simple, and the whole arrangement described by that sentence is composite. Already for this setup, we are invited to consider other possibilities: couldn't we consider each square as composed of two smaller rectangles, or of color and shape, for example?

In §64, Wittgenstein refers back to this game and imagines a variation[7]. Consider interpretation schema 2 (Figure 3.1c) and the letters 'U', 'V', and 'X' as representing, respectively, a red square above a green square above a red square, a red square above a green square above a white square, and a black square above a green square above a white square. In this variation, one could say that the sentence 'UVX' also

[6] I look more closely into these passages in Section 5.2.

[7] The following is not exactly what Wittgenstein describes, but it captures, I believe, the same point.

describes the arrangement in Figure 3.1a. Is then, the sentence ‘RRBGGGRWW’ a more fundamental analyzed form of ‘UVX’? Could we replace the language-game that uses interpretation schema 2 with the one described in §48? Wittgenstein’s answer is a negative one: “It is just a *different* language-game; even though it is related to (48).” (2009, §64b, p. 35e) Based on previous remarks, one can see at least two reasons underwriting this reply. First, someone who knew how to play the language-game of §64, by using letters to describe columns could see each column as a unit with a special character, just like we see the French tricolor as more than just an arbitrary arrangement of three colors. Second, in order to understand the sentence ‘RRBGGGRWW’ one would need to know how to play the language-game of §48. This includes knowing that the sentence is (in the context of that game) composite, knowing the colors associated with each of the letters, and knowing the order represented by the interpretation schema 1. Thus, explaining ‘UVX’ as ‘RRBGGGRWW’ is neither necessarily meaning-preserving nor self-contained. One can summarize the take-home message by the following remark:

Remark 3.2.1 *An explanation is always anchored in a language-game, and its value is relative to that language-game.*

There are also remarks that hint at further issues with some expectations we might have regarding explanations that are linked with the aforementioned regressive aspect of analysis. One of the first comes as early as §1d. Wittgenstein sets up a hypothetical situation where a shopkeeper is given a paper with the words ‘five red apples’ written on it and describes the actions performed for each word. The following dialogue between narrator and interlocutor follows:

> “But how does he know where and how he is to look up the word ‘red’ and what he is to do with the word ‘five’?” — Well, I assume that he *acts* as I have described. Explanations come to an end somewhere. – But what is the meaning of the word “five”? – No such thing was in question here, only how the word “five” is used. (2009, §1d, p. 6e)

We can see the interlocutor’s questions as characteristic of a regressive urge. He is trying to understand the words ‘red’ and ‘five’ in this context, and for that he seems to be looking for some apodictic foundation. The narrator is going one level deep by describing the behavior of the shopkeeper when being handed the paper with the words on them, but the interlocutor wants to go deeper. The narrator deflects this urge twice, each being characteristic of Wittgenstein’s thought. I discuss some implications of the second deflection in Section 4.2. As for the first deflection

("Explanations come to an end somewhere"), it seems to me to be revealing of an attitude towards philosophical method that surfaces in other sections as well, particularly clearly in §29 and §87.

The problem with regression is clear and far from new. Whenever we provide an explanation of the meaning of a word, this is done in terms of other words. What keeps us from asking for an explanation of the meaning of the words contained in that explanation? And subsequently of an explanation of those? A foundationalist might argue that such a regress cannot go on indefinitely, otherwise no one could ever understand an explanation, so one must assume the existence of final self-sustained explanations. Wittgenstein, however, takes a different approach:

> These questions would not even come to an end when we got down to words like "red", "dark", "sweet". – "But then how does an explanation help me to understand if, after all, it is not the final one? In that case the explanation is never completed; so I still don't understand what he means, and never shall!" – As though an explanation, as it were, hung in the air unless supported by another one. Whereas an explanation may indeed rest on another one that has been given, but none stands in need of another – unless we require it to avoid a misunderstanding. One might say: an explanation serves to remove or to prevent a misunderstanding — one, that is, that would arise if not for the explanation, but not every misunderstanding that I can imagine.
> [. . .]
> The signpost is in order – if, under normal circumstances, it fulfils its purpose. (2009, §87, p. 45e)

Words (or other linguistic entities) do not require, in and of themselves, for their meaning to be explained. *We*, agents making use of words, may require explanations whenever we encounter or expect a misunderstanding.

If we place a signpost at a crossroads where one road leads to village A and the other to village B, indicating which leads where, and we subsequently find that people end up in the village they intended to go to, we do not need to add another signpost explaining how the first is to be interpreted, or a third one explaining the second. If, however, people following the road usually get lost, we might indeed add a further explanation. But that does not mean that, just because this possibility exists, a second signpost is always needed. In the context of analysis, this position is not exactly against the regressive aspect altogether. An explanation may rest on another, and may be helpful to resolve a misunderstanding. Problems arise when one goes looking for an explanation without a misunderstanding that needs to be resolved, without a practical purpose. Explanations do come to an end at some point, but that point is dependent on the purpose.

Not only should the need for an explanation be dependent on a practical purpose, but one should also bear in mind that both the explanation and the purpose are situated in a context. In §§28-29, Wittgenstein discusses an example of how one could explain the meaning of the word 'two' by pointing at a pair of nuts. The issues of regression quickly arise here too: if, in order to avoid a misunderstanding, one would accompany the gesture by the phrase "This *number* is called 'two'.", wouldn't one potentially also need to explain the word 'number'? And again to explain that explanation by means of other words, continuing this exercise *ad infinitum*? This concern is, however, countered by the following remarks:

> Whether the word "number" is necessary in an ostensive definition of "two" depends on whether without this word the other person takes the definition otherwise than I wish. And that will depend on the circumstances under which it is given, and on the person I give it to. (2009, §29, p. 18e)

The dissolution of the issue is here similar to that in §87: there is no need to provide further explanations unless the ostensive definition is taken otherwise than is intended, *i.e.* unless it fails to fulfill its purpose. The important addition to this is the reminder that whether an explanation succeeds or fails also depends heavily on context. This implies that there is no foolproof *a priori* criterion for a good explanation and "[a]ny explanation can be misunderstood" (2009, §28b, p. 17e). One can imagine a number of circumstances having an impact on this (*e.g.* whether one is pointing at two nuts isolated on a table, or whether they are far away and next to another group of three nuts), as well as personal idiosyncrasies that might make the listener fail to understand the explanation under the same circumstances (the most obvious being whether or not she knows the word 'number'). These observations fit well with Wittgenstein's picture of language as a practice (see Section 4.1) by drawing attention to contingent circumstances and the involvement of agents. An explanation is thus dependent on a situated purpose:

Remark 3.2.2 *Explanations should only be pursued when driven by, and in the context of, a situated purpose; their success cannot be guaranteed independently thereof.*

This conception of explanation demonstrates an attitude that is very different to the one criticized in Section 3.1. Explanations of meaning are not to be judged on their intrinsic properties, but rather on whether they fulfill their purpose in the situated context where they are given. They can be ambiguous (§28b), vague (§71, §88), and incomplete (§87), as long as they are useful. They can also be misunderstood, any explanation can (§28b, §71) and there is, therefore, no absolute objective criterion for their success. Given how explanations are anchored in language-games (Remark 3.2.1, p. 67), and that new ones are continuously invented and other ones get forgotten (§23), explanations are inevitably provisional; they are relevant while someone knows and is interested in playing the language-games that underwrite them. This attitude abandons the ideals of exactness, objectivity, and atemporality by pivoting on the primacy of contingent and situated purpose. It aims to embrace the heterogeneous and dynamic nature of meaning, and reject the urge to look for *a priori* apodictic answers to requests for the meaning of linguistic expression.

These considerations have a bearing on philosophical method for a number of reasons. First, as was discussed in Section 3.1, philosophical inquiry starts, in Wittgenstein's eyes, from questions of meaning, for which a certain preconception of language leads the philosopher to look for exact, objective, atemporal answers:

> We ask: "*What is* language?", "*What is* a proposition?" And the answer to these questions is to be given once for all, and independently of any future experience. (2009, §92b, p. 48e)

Philosophical methods can be seen as attempts to systematize how one goes about addressing these questions. Philosophical analysis, if conceived in terms of decomposition, involves explaining a concept in terms of other, supposedly simpler concepts. Explication, phenomenology, connective analysis, ordinary language philosophy, and other methods might prescribe different approaches, but, according to Wittgenstein, they are ultimately trying to understand the use of words, *i.e.* philosophical problems are not empirical (§109), and the inquiry is rather a grammatical one (§90). Following a philosophical method can thus be seen as providing a sophisticated and more or less systematic explanation of meaning, but an explanation of meaning nonetheless. Therefore, whatever virtues Wittgenstein exalts in explanations of meaning should be relevant for philosophical methods as well. Second, it is no coincidence that §§87-88, which heavily feature remarks on explanation, are followed by the most markedly metaphilosophical part of the *Philosophical Investigations.* Not only that, but they are even connected by the following passage: "With these considerations we find ourselves facing the problem: In what way is logic something sublime?" (2009, §89a, p. 46e) The immediately preceding considerations in §§87-88 defend a particular conception of explanations of meaning. That these seem to lead to thoughts on logic (broadly construed), analysis, and method in philosophy, is further evidence that explanation and these issues are strongly connected.

Wittgenstein's attitude towards explanations of meaning has pragmatist undertones. As I argued in this section, it is highly contextualist: explanations are portrayed as always tied to particular language-games, and their interpretation and success as depending on the circumstances surrounding their use, including the person the explanation is given to. Furthermore, no language-game is necessarily privileged over another, some are simply different even if they can be related. If we see philosophical methods in a similar light, we can interpret this as supporting the following claim: if the force of a method depends on the context, and there is no privileged language-game, there can be no single optimal way to approach every philosophical question. This aligns well with the picture of Wittgenstein as committed to methodological pluralism, as mentioned at the end of Section 3.1.

This pluralism is not full-blown relativism, on the contrary. First, philosophical language-games constitute the context in which each method is evaluated. Attempts to use conceptual analysis or formal logic in a conference on the philosophy of Jacques Derrida are likely to be as denounced as arguments based on deconstructivism in a meeting on analytic metaphysics, whereas the former methods would probably not raise any eyebrows in the latter gathering, and vice-versa. Importantly, as language-games are social practices (see Chapter 4), the success of each method further depends on the people involved and the way they act and interact with each other. Second, Wittgenstein has his own criteria of approval. As I argued in Section 3.1, they apply not to particular methods, but to the attitude one has towards them. Thus, he would likely reject the use of conceptual analysis if it was presented as finding some objective final analyzed form of a linguistic expression, but might not have any qualms with the method being used as a provisional tool to dispel some perceived misunderstanding.

A common objection to pluralism could be raised. Isn't the claim that there is no objectively optimal method itself in need of an Archimedean point to derive its force from? What are the basis for Wittgenstein's criteria and picture of language, and why should one accept them? Such concerns reflect the very attitude being criticized. They demand for ultimate foundations, where the claim is precisely that these foundations do not exist. Reflecting on the remarks on explanations of meaning discussed above can illuminate a possible reply. Wittgenstein's position with regards to explanations of meaning, I argued in this section, is an anti-foundationalist one: there are no ultimate self-explanatory explanations on top of which all the others rest. What can someone who defends such a position reply when continuously pressed for justifications?

> Once I have exhausted the justifications, I have reached bedrock, and

> my spade is turned. Then I am inclined to say: "This is simply what I do" (2009, §217b, p. 91e)

Explanations come to an end somewhere. But this end is not a self-sustained foundation or an apodictic truth. It is simply the point beyond which communication breaks down, and disagreements persist. An anti-foundationalist cannot give an irrefutable argument that there are no irrefutable arguments. Nor can someone who claims that there is no single optimal method provide an infallible justification for that claim. But shouldn't the burden of proof be on the other side? Exploring this issue in detail is beyond the scope of this paper, as it would require delving into a thorough discussion of the old and unresolved problems typically associated with relativism. And Wittgenstein himself does not provide a critical examination of his position. The important thing to retain is Wittgenstein's methodological pluralism, which we can summarize in the following remark.

Remark 3.2.3 *There is no single valid philosophical method, but there are many useful ones.*

In Section 3.1, I argued that Wittgenstein's critical points found in §§89-133 are directed not against particular methods, but against a certain attitude towards them, namely the tendency to see them as means to discover atemporal apodictic truths. Establishing a parallel between explanations of meaning and philosophical methods allows us to see an alternative. If one employs a philosophical method with a practical purpose in mind, with full awareness of its locality (anchored in particular language-games, situated in a context), one is able to relinquish the craving for exactness in favor of fulfilling the purpose at hand, in the context at hand. That Wittgenstein's remarks are not in defense of one particular method, but of a general attitude towards them, is further supported by the methodological pluralism both advocated for explicitly (*e.g.* in §133d) and embodied in his practice. This, I believe, leaves plenty of room for systematicity (as method-following rather than system-building) in philosophy. What is crucial is to maintain a pragmatist attitude towards the methods one uses.

When it comes to the study of natural language, it is easy to interpret Wittgenstein's rejection of 'logic' and the ideal of exactness, and his appeal to go "[b]ack to the rough ground" (2009, §107, p. 51e), as incompatible with any kind of formal approach. Even Richard Rorty defends that "if one adopts the point of view of Wittgenstein's *Philosophical Investigations*, there can be no such thing as a 'systematic theory of meaning for a language'" (1991, p. 57). I want to argue that signaling games, despite being a formal approach to the study of meaning, can be seen as retaining the pragmatist attitude that is dear to Wittgenstein, while nevertheless being very systematic. It is important, however, that we do not see them as a theory of meaning. They are a framework, a toolbox so to speak, that constitutes a method in the same way as Wittgenstein's artificial scenarios of language use that we encounter throughout the book. With that in mind, it is important to reflect on what is said about these toy language-games, as I shall here call them.

3.3 The use of toy language-games as method

One can separate Wittgenstein's use of the term 'language-game' into two kinds. One refers to the various real ways in which people use language and the activities these are interwoven with. The best list of examples of such activities can be found in §23, and includes things like describing an object by its appearance, making up or reading a story, acting in a play, or telling a joke. These are scenarios of language use that we can identify as part of our existing practices. Another sense in which the term 'language-game' is used refers to admittedly artificial scenarios of language use created by Wittgenstein in the context of discussions of specific issues. The best examples are the often-called builders' language, introduced in §2 and extended in §8, §15, and §86, the language of colored squares, introduced in §48 and further discussed in §64, and the beetle-in-the-box thought experiment of §293. Baker and Hacker (1980) distinguish these two senses as *natural* and *invented* language-games.

I prefer to call the latter *toy* language-games, because they resemble the kind of toy models often used in science and engineering (see Frigg and Hartmann, 2018). Both are deliberately simple descriptions of self-contained hypothetical systems that are supposed to capture the relevant elements that allow one to study particular aspects of larger natural systems. They do not purport to be an accurate representation of reality, but rather to allow one to play with them in order to better understand more complex phenomena. Both are cases of thought experiments: the author creates a specific hypothetical scenario, introducing a number of entities and establishing the relations between those entities, and then explores the plausibility of intuitions about the scenario, from how it should be interpreted in light of a certain theory, to how it would play out if left to the specified devices.

In order to think of the use of toy language-games as a method, and before one considers questions of systematicity, it is important to reflect on how one can or should conceive of them according to Wittgenstein. Namely, this includes important questions regarding what these models are, how they can or should be used, and how one could develop them further. When it comes to toy language-games, Wittgenstein gives us some clues on his own attitude towards them. First, we are given two ways of conceiving of them. With respect to the builders' language, we are told that we can imagine it as a "complete primitive language" (2009, §2, p. 6e), *i.e.* a system of communication used by the builder and the assistant, or even by a whole tribe of people engaged in such activities (§6). Another option is to see the builders' language as "one of those games by means of which children learn their native language" (2009, §7, p. 8e). Even though a child eventually learns a more complex language, the process of language acquisition starts with interactions that are similar to toy language-games in many ways (see Ratner and Bruner, 1978 and Tomasello, 2008, §4.4). In both readings, toy language-games are thus to be imagined as complete, in that they enable conducting a certain activity without the need for anything else, and isolated, in that they do not necessarily interact with other language-games (in particular because, both in the case of being a primitive language or of being a game by which a child learns to talk, there are no other language-games to be interacted with).

Remark 3.3.1 *Toy language-games are complete and self-contained hypothetical scenarios of language use.*

Studying toy language-games enables abstracting away from the complexities of natural language. Whereas natural language-games are diverse, ever-changing, and interwoven with each other, each toy language-game is imagined as complete and is typically discussed in isolation. The purpose is to gain clarity:

> If one looks at the example in §1, one can perhaps get an idea of how much the general concept of the meaning of a word surrounds the working of language with a haze which makes clear vision impossible. – It disperses the fog if we study the phenomena of language in primitive kinds of use in which one can clearly survey the purpose and functioning of the words. (2009, §5, p. 7e)

If one thinks as language as a whole, and aims for a general theory of meaning, as philosophers of language often do, clear vision is, according to Wittgenstein, impossible. However, if one sees language as composed of smaller units, it becomes legitimate to focus on one specific kind of use. In particular, one can do so by creating a toy language-game as a simplified example of that kind of use. This can allow one to get a clearer overview of the elements at play. Toy language-games are tailored to specific questions. The builders' language contains objects in order to motivate reflections on meaning as reference to the external world, the colored squares game has complex sentences to illustrate problems of logical atomism, and the beetle-in-the-box imagines a super-private entity to raise issues with psychologism (more on all of these in Chapter 5). It would be very difficult to talk about compositionality in the context of the builders' language, but this does not invalidate the usefulness of that particular toy language-game to motivate discussions about the kind of use it illustrates.

Such an approach of independently studying particular kinds of use can raise concerns that one could fail to see the forest for the trees. Wittgenstein's position, however, is that language does not have an essence (§92) or a formal unity (§108). Language-games are interrelated, but there is nothing necessarily common to *all* of them (§§65-67). If what is called 'language' is a family concept in this sense, it is better to try to understand it in the same way one understands other family concepts, like 'game' or 'number': by getting better acquainted with various examples, building others on the analogy of these, and so forth (§75). This is exactly what Wittgenstein does with toy language-games: they "stand there as *objects of comparison* which, through similarities and dissimilarities, are meant to throw light on features of our language" (2009, §130, p. 56e). Rather than starting with a theory of meaning built on the assumption of the existence of common features to all that we call language, and then interpreting the particular instances in terms of that theory, Wittgenstein is suggesting one should start by studying individual kinds of use and develop further understanding of language by contrasting several such instances with each other. It is an approach that is bottom-up, rather than top-down.

Remark 3.3.2 *The most productive way to study language is to independently study specific kinds of use.*

Another important difference between toy language-games and other systematic approaches to studying language is one of attitude towards the method. In line with what was pointed out in Sections 3.1 and 3.2, Wittgenstein says the following:

> Our clear and simple language-games are not preliminary studies for a future regimentation of language – as it were, first approximations, ignoring friction and air resistance. (2009, §130, p. 56e)
>
> For we can avoid unfairness or vacuity in our assertions only by presenting the model as what it is, as an object of comparison – as a sort of yardstick; not as a preconception to which reality *must* correspond. (The dogmatism into which we fall so easily in doing philosophy.) (2009, §131, p. 56e)

These are warnings directed at those who might feel certain urges towards toy language-games. In particular, toy language-games are not to be taken as simplifications that can eventually be generalized into a theory of meaning. Let alone are they aimed at capturing any essential property of language or meaning. This rejection stems from the picture of language as a family: if there is nothing necessarily common to *all* language-games, there is no essence to what language is, and thus no general theory of meaning to be constructed.

Like explanations of meaning, each toy language-game is best seen as local to the particular aspect of language use it exemplifies. That the builders' language is not appropriate to discuss compositionality is neither a flaw of that toy language-game nor a sign that it needs to be extended. That toy language-game has a situated purpose (Remark 3.2.2, p. 69), namely to help someone better understand a kind of referential use of language, and it is in order as long as it fulfills that purpose. This is not to say that it *cannot* be extended, Wittgenstein does so a number of times (§8, §15, §86), but only that it does not have to. One must resist the craving for universality (Remark 3.1.3, p. 61). Toy language-games should be seen and presented as models, objects of comparison, or yardsticks, *i.e.* as tools to further our understanding of particular kinds of use of words and sentences, and thus indirectly of the workings of language. The relevant lesson can be summarized in the following remark:

Remark 3.3.3 *Toy language-games are not to be generalized into a theory of meaning.*

One might find the use of toy language-games problematic given that the notion of language-game, their natural counterpart, is never very clearly defined. The remarks that follow §65 attempt to address this issue. As already mentioned, Wittgenstein rejects the need for an essentialist definition (in §67 and other places). In reaction to this, the interlocutor suggests a disjunctive characterization: one could define a family concept, like 'game' or 'number', as "the logical sum of corresponding sub-concepts." (2009, §68a, p. 37e) This, the reply argues in the following remarks, would be to draw a rigid boundary where no boundary is needed. We use concepts like 'game' and 'number' successfully without making this kind of exact definitions. The same should be possible with the notion of a language-game.

The concept of a language-game is thus admittedly vague, based on rough allusions, metaphors, and enumerations of examples. There is neither a statement of necessary and sufficient conditions nor a formal procedure to decide when something qualifies as a language-game or not. This is explicitly by design. It is tied not only to the picture of language as heterogeneous and dynamic, but also to a deeply pragmatist stance. When pressured by the interlocutor regarding the possibility of making a more exact definition of 'game', Wittgenstein retorts:

> To repeat, we can draw a boundary – for a special purpose. Does it take this to make the concept usable? Not at all! Except perhaps for that special purpose. (2009, §69, p. 37e)

Regarding the Fregean idea that a blurred concept is no concept at all, he further says: "This presumably means that we can't do anything with it." (2009, §71, p. 38e) The idea Wittgenstein is setting up as problematic is that a concept needs an exact definition in order to be used.

Both his rejections are thus anchored in considerations of usefulness: what matters is not whether a definition of a notion like language-game is exact (whatever that may be), but whether or not one can do something with it. For example, if I want to take a picture of you and tell you to "stay roughly here" (2009, §71, p. 38e), precise GPS coordinates of where to stand are typically neither necessary nor even desired; what matters is that I fulfill the purpose of taking a picture in a way that satisfies us both. The need for more or less precision is tied to the particular purpose we have in mind, and to the extent to which the words used serve that purpose. The discussion in these sections leads to remarks on vagueness and rule-following that will be addressed in later sections. For now, the important thing to retain is Wittgenstein's explicit acknowledgment and acceptance of the loose definition of the notion of language-game:

Remark 3.3.4 *It is acceptable to define language-game by metaphors and examples; a stricter definition should only be advanced when needed for a special purpose.*

These remarks give us an overview of how Wittgenstein conceives of toy language-games and their intended use. The way he himself uses them is already somewhat systematic, in the sense defined in the beginning of this chapter. He himself talks about applying "the method of §2" (2009, §48a, p. 27e) and uses it in several moments throughout the *Philosophical Investigations.* The method is described by Stern (2004, pp. 10-15) as consisting of three stages: first, there is a characterization of a certain philosophical idea; second, a toy language-game is imagined where this idea would apparently apply quite well; third, observations are presented to make us see that even in this ideal scenario the idea is much more limited than it initially appears. Toy language-games are thus a tool to explore the assumptions of philosophical ideas about meaning by laying them bare in a concrete example and reflecting on how such a scenario would play out. If it fails to match our expectations we have exposed a problem with the philosophical idea. Concomitantly, they can also serve to illustrate and motivate a different perspective on meaning (see Chapter 4). This is patent in, for example, how Wittgenstein uses the builders' language to both raise issues with referentialism and promote the shift of attention from reference to use (see Section 4.2). In the following section, I reflect on the resemblance between toy language-games and signaling game model, and how well the way they are used aligns with the remarks on methodology brought forward in this chapter so far.

3.4 Systematizing toy language-games

When comparing toy language-games and signaling game models, one can start by considering their structural similarities. In Section 2.2, I laid out the typical components of signaling game models. Toy language-games can also be seen as put together by combining a number of various elements. Baker and Hacker (1980, pp. 54–55) identify seven. To their list, I add two additional ones[8], agents and objects, and remove one, completeness[9]. I thus consider the following elements as the relevant components of toy language-games:

Agents* I start with one element that Baker and Hacker fail to mention, but that I believe is important to include, especially when considering Wittgenstein's picture of language as practice (see Chapter 4): the presence of agents. For example, in the builder's language of §2 there is a builder and an assistant, in the colored squares language-game there is an individual A describing colored squares to another individual B (see §49), and in the beetle-in-the-box game (§293) there are a number of people using the word 'beetle'. Discussions of meaning never present language in isolation, toy language-games are always described with reference to agents using language in the context of a practice;

Vocabulary Consists of a set of linguistic entities (*e.g.* words, sentences) available to the agents. It is typically a finite set, like 'block', 'pillar', 'slab', and 'beam' (§2), but it can be a vaguely defined potentially infinite set, like numbers (§21) or series of signs (§143);

[8] I mark these with '*' for the sake of clarity.

[9] Completeness is not an element of a toy language-game in the sense that things like agents, or a vocabulary, are. It is at best, following Remark 3.3.1 (p. 74), simply a desired property of them.

Instruments Includes additional elements that could be said to form part of the vocabulary but are usually not considered linguistic, such as gestures (*e.g.* §28), samples (*e.g.* §50), or pictorial representations (*e.g.* §291);

Objects* While instruments and vocabulary are used by the agents as signs, real objects are sometimes also part of toy language-games in a different role. Examples are the building stones in the builders' language (§2), the sword Nothung in §39 and §44, or the chair in §80. While these elements of the external world do not constitute the meaning of elements in the vocabulary, they are relevant to the use that is given to those elements by the agents involved;

Activity Some toy language-games specify actions that agents can perform. The best example is the builder's language where the assistant chooses which stone to bring to the builder. Toy language-games like the colored squares and the beetle-in-the-box do not talk about physical actions, but they are still, as discussed in Section 4.1, always considered as kinds of activities;

Purpose An additional ingredient in discussions of meaning is the purpose linguistic entities are being used for. Toy language-games thus often characterize the objectives of the agents involved. For example, in the builder's language it is relevant to the meaning of the call "Slab!" if the builder wants to simply evoke a mental picture in the assistant, or if he wants the assistant to perform a certain action. It should be noted that not all toy language-games characterize this in detail;

Context There are circumstances beyond the explicit elements available to the agents (vocabulary, instruments, objects, and actions) that can additionally influence the use of language. Baker and Hacker describe these as the "presuppositions of meaning" (1980, p. 54), and include things like general features of the natural world*, e.g.* that chairs do not usually disappear and reappear suddenly from sight (§80), and general features of human beings that enable us to begin to interpret an unknown language (§206);

Learning processes Discussions of toy language-games sometimes include not only the playing of the game, but also speculations about the processes required for an agent to learn how to play the game in the first place. The most detailed examples include the learning of the builders' language in §§6-9 and the learning of the number series language-game in §§143-155.

Baker and Hacker (1980, p. 55) also talk about completeness, but those considerations are not exactly about any element of toy language-games, but rather about Wittgenstein's attitude towards them.

With these elements in mind, and recalling the typical anatomy of a signaling game (see Section 2.2), it is possible to make a point by point structural comparison. As in toy language-games, considerations about agents are also always present in signaling game models. There is no game without individuals to play it, and their characterization has repercussions to other elements in the model as well. Researchers thus tend to represent them in much more detail than Wittgenstein does. This includes making choices about their level of rationality, whether one is dealing with two (or more) agents or populations thereof, what kind of strategies should be used to represent them and how they should be updated, among other things. Vocabulary in toy language-games corresponds clearly to message space in the signaling games framework. Messages can represent words or sentences used by the agents in communication. As mentioned in Section 2.2, message spaces can be finite or infinite, but also dynamic, which is an aspect in which signaling games can be seen as going beyond toy language-games.

Instruments do not have a clear correspondence in signaling games. Whether and how they can be incorporated depends on their role. Things like gestures, for example, can be easily seen as messages. The latter need not be linguistic entities; they simply stand for anything that the sender can produce and the receiver can identify. Note that there is nothing ontologically special about what is exchanged when we say we are using words or sentences. Ultimately, what is produced by the sender is either a sound or an action that results in a visual pattern (like for example ink marks on paper). A gesture fits this description just as well, and can thus be represented in a signaling game model by a message. In fact, Lewis originally describes signals as kinds of actions (1969, p. 122).

Other instruments, like for example samples, are more difficult to fit in the picture. Wittgenstein talks about them as means of representation (§50). They are objects used to make certain types of statements, like the standard meter can serve as a reference point for statements about length. They could be considered as part of the message space, but would have to be assigned a special role. This could perhaps be achieved in a model with complex messages but, to the best of my knowledge, this has not been attempted in the literature yet.

Objects in toy language-games can be seen as related to either states or actions in signaling games. Pillars, slabs, swords and chairs are the kind of things that can trigger senders to produce particular messages in response to which they expect certain actions to be performed by receivers. Note that the absence of objects can easily be seen as having the same effect. The need for a pillar, rather than its presence, can trigger the sender to produce a signal in order to obtain the missing object. States and actions are more than mere objects, but objects can often (though not necessarily) be related to these two components of signaling games. What Baker and Hacker call activity can also be related to states and actions. This is clearly the case for the receiver but, as discussed above, messages can also be interpreted as actions performed by the sender. Again, what is relevant is that whatever the sender does for a given state is accessible to the receiver, and that the latter makes a choice based on it that has implications for both, but the framework is certainly motivated with activity in mind.

The implications of agents' choices are captured, in a signaling game model, by the utility function. This can be seen as representing what in toy language-games is called purpose by Baker and Hacker. As mentioned in Chapter 2, what the utility function represents has at least three possible readings in the signaling games literature. If a signaling game model is interpreted in terms of classic game theory, like the ones originally proposed by Lewis (1969), one can see utility as encapsulating the preferences of the rational agents involved in the game. This aligns well with how purpose is described by Wittgenstein in the context of toy language-games. If a model is analyzed using learning dynamics, utility is probably best seen as representing the way the agents are externally reinforced. Here, one can still see utility as related to purpose. Utility would no longer represent the purpose of the agents playing the game, but rather that of some potential tutor. Note that, however, just because there is learning and reinforcement that does not imply that there is a conscious purposeful tutor. One can learn from acting in the world and have one's behavior reinforced simply because some choices yield positive and others negative results. Finally, if a signaling game model is understood in terms of populations of agents driven by evolutionary forces, seeing utility as purpose would amount to a teleological take on evolution, an interpretation which is controversial, to say the least (see Allen and Neal, 2019). Selection is best seen as a mechanistic process. In the context of that perspective, utility represents fitness, which can be seen abstractly as likelihood of replication, like survival and reproduction in biological evolution, or propagation by imitation in cultural evolution. Therefore, in signaling game models aiming to capture these types of processes, like the ones proposed by Skyrms (1996), the notion of purpose present in some toy language-games plays no role.

The element of context is somewhat vague and ill-defined, hence it is difficult to make a clear comparison. The specification of prior probabilities for states in the state space is a small component of the signaling games framework that could be considered as forming part of context. Priors can be interpreted either as the true distribution of states, in which case they could be seen as general features of the world, or as how likely agents think each state is, in which case they could be seen as forming part of their presuppositions. In general, however, contextual features will typically end up embedded in different components of each model. This can include how the state, message, or action spaces are structured, how agents' abilities and commonalities are reflected in the dynamics, what defines success or failure as captured by the utility function, and so on. Ultimately, much like in toy language-games, these features will vary strongly from model to model.

Learning processes turn up in a lot of the different ways of signaling game models can be analyzed. As will be discussed in Section 4.3, if a model is seen as a single shot game, there is no room to capture any kind of dynamic adaptation, including learning. In the context of repeated games, however, there are a number of options to represent different types of learning processes. These range from learning by imitation, as in the interpretation of the replicator dynamics as cultural evolution, to reinforcement learning, a more behavioristically inspired model, to other processes available to agents with higher levels of rationality, like the best-response dynamics. If one is interested in learning processes, the signaling games framework is thus well suited to take them into account. Additionally, there are also other ways to interpret and analyze models, for example in terms of rational deliberation or biological evolution.

In order to illustrate these connections between the elements of toy language-games and the components of the signaling games framework, it can help to consider how one could build a concrete model for a particular example. Take the builder's language of §2. The agents in the toy language-game are the builder and the assistant. Given their roles, one could think of the former as the sender, and the latter as the receiver. Their possible behavior (the sender choosing one word to use, and the receiver choosing one building stone to hand over) could be represented in terms of simple pure strategies. The message space could be constituted by a discrete finite set with a message per word in the vocabulary ('slab', 'pillar', etc). States would have to represent what triggers the builder to use those words. In the context of Wittgenstein's story, that is the need for a particular building stone. Thus, there could be a state for being in need of a slab, another for being in need of a pillar, and so on. Conversely, the actions available to the receiver include handing over a slab to the builder, handing over a pillar, and so forth. The utility function would encode the purpose of the game, which presumably is for both sender and receiver to satisfy the sender's need for a particular building stone. This could be formalized by a function that trivially assigns a utility of 1 to the combinations of actions that satisfy the states (*e.g.* the assistant hands over a slab when the builder is in need of a slab) and 0 to other combinations. This is a simple Lewis signaling game (as defined in Chapter 2) that could be analyzed in terms of learning processes, much like it is done for the builder's language in §§6-7. I did not specify any instruments or additional context, and the existence of objects is only implicit, but not all elements are relevant for all games.

It seems clear that the signaling games framework motivates representing scenarios of language use in a way that is structurally similar to Wittgenstein's toy language-games. However, there also dissimilarities that need to be noted. One is that toy language-games are mostly used to support remarks of a negative nature. Stern (2004) emphasizes this aspect of the 'method of §2':

> This three-stage argument scheme suggests a more general recipe for unsettling philosophical preconceptions. First, describe a case the preconception fits as well as possible, [...] then change just enough about the case in question [...] so that we run up against the limitations of the preconception. (2004, p. 11)

According to his interpretation, toy language-games are used with the purpose of supporting the dissolution of a philosophical preconception. This is indeed the main angle of most uses of the method. For example, the builder's language supports the argument that meaning is more than reference to external entities, the colored squares game is used to reveal problems with ideas of logical atomism and compositionality, the beetle-in-the-box experiment can be seen as illustrating the incongruence of psychologistic theories of meaning, and other toy language-games are used to dismantle some intuitions about rules and rule-following.[10] This should come as no surprise given the many passages in the *Philosophical* Investigations, urging philosophers to focus on clarificatory tasks (see again Section 3.1).

Signaling game models, unlike toy language-games, are usually created to support positive hypotheses about language and meaning, as can be seen from the literature discussed in Chapter 2. *Prima facie*, they seem to therefore fail to align well with Wittgenstein's later philosophy when it comes to explicit considerations about method. It might very well be the case that he would not have endorsed their use for this very reason. However, as pointed out earlier, there is a certain conflict

[10] See Chapter 5 for more details on the first three cases, and Chapter 6 for the toy language-games regarding rules and rule-following.

between metaphilosophy and practice in the *Philosophical Investigations*. Wittgenstein's position hinges on a particular picture of language (see Chapter 4) which is difficult not to see as a set of hypotheses. That language is heterogeneous and dynamic, that one should see it as a practice, that it is constituted of interrelated language-games, that one should look at use in order to better understand meaning, these are all suggestions that form a positive view on how language is like. Despite being presented as platitudes, they are not obvious, not shared by everyone, and are general enough to need further evidence and substantiation.

This picture of language (what Plant, 2004 calls Wittgenstein's minimal dogmatism), is to some extent motivated for in the context of discussions that make use of toy language-games. The builder's language supports not only the negative claim that learning reference is not sufficient to learn the meaning of a word, but also the positive claim that ostensive teaching "together with a particular kind of instruction" (2009, §6, p. 8e) can bring about understanding. The beetle-in-the-box experiment implies that, as long as the word 'beetle' has a use in the language-game, it has a meaning, even when the box is empty (§293). The colored squares example is used to illustrate how two language-games can be different but related (§64). And more examples could be pointed out. To the extent that the aforementioned picture of language constitutes a positive view, and that toy language-games are sometimes used to promote it, I think that making use of the signaling games framework to argue for further positive hypotheses about language and meaning can be in line with Wittgenstein's practice.

If one accepts this as valid method in philosophy, one should still take heed of Wittgenstein's remarks on how philosophy often goes astray. Philosophical questions are often created by a picture of language anchored in correspondence. A failure to see this, combined with a craving for exactness, can lead one to misjudge the way one should approach these kinds of questions (Remark 3.1.1, p. 57). This can lead one to look for hidden essences or universal atemporal answers (Remark 3.1.3, p. 61) and this has an important impact on which methods one thinks are suitable to approach the questions. If, however, one adopts a different picture of language and recognizes that linguistic expressions do not necessarily have single analyzed forms (3.1.2), one can stop looking for exact definitions, necessary and sufficient conditions, or necessary truths as ends in themselves. In Section 3.1 I argued that Wittgenstein's negative views are not against particular methods, but against a certain attitude towards them. It is important to keep the aforementioned remarks in mind when making use of the signaling games framework as well.

This brings me to another aspect in which toy language-games and the signaling games framework differ. Whereas the former are purely verbal models, setting up the elements and their interrelations solely in terms of natural language, the latter make use of mathematical language and often of computer simulations. As such, signaling game models are represented and analyzed with a much higher degree of precision than toy language-games. This can be seen as succumbing to the craving for exactness Wittgenstein warned against (see Remark 3.1.1, p. 57). Note, however, that the problem with exactness is tied to a picture of meaning as correspondence. Philosophers are lead into confusion when they assume elements of discourse necessarily map to real entities which must therefore be possible to characterize precisely.

Signaling models are admittedly artificial and abstract; their elements do not correspond to real entities, and the level of precision they are described in and analyzed with does not therefore need to be projected onto reality. Precision is, in the signaling games framework, merely instrumental. Taking into account what was said in Section 3.2 as applicable to both toy language-games and signaling games alike, additionally motivates related considerations. In particular, one should always keep in mind that these methods are contingent, both to the particular context in which they are being employed (Remark 3.2.1, p. 67), as well as to the purpose at hand (Remark 3.2.2, p. 69). As such, no use of these methods should purport to be universal, in the sense of capturing an essential property of everything that we call language, or claim to necessarily represent reality. As long as all of this is kept in mind, the use of mathematical language and computer simulations does not need to lead to the same problems that the philosophical endeavors portrayed in the context of Remark 3.1.1 (p. 57) do.

These formal tools make signaling games more systematic than toy language-games. Going beyond the aforementioned structural similarities and comparisons in terms of methodological aims, we can see additional resemblances by exploring how both toy language-games and signaling game models can be seen as instances of thought experiments. Thought experiments have been used in both science and philosophy for as long as these practices exist (J. R. Brown, 1991; Rescher, 2005). There are ongoing debates on what thought experiments are and what epistemic powers they actually have. It is beyond the scope of this thesis to engage with that debate[11]. I concur with Nersessian (1992) in seeing thought experiments as a form of "simulative model-based reasoning" where someone reasons "by manipulating mental models of the situation depicted in the thought experimental narrative." (1992, pp. 291-292)

It should be clear by everything said so far that both the use of toy language-

[11] See J. R. Brown and Fehige (2017) for an overview of the issues.

games and of the signaling games framework can be characterized in these terms. Toy language-games are complete and self-contained hypothetical scenarios of language use that certainly fit the description of thought experiments (Remark 3.3.1, p. 74). Furthermore, the method of §2 fits well into what El Skaf and Imbert (2013) call the CUI pattern of inquiry: *construction* of a scenario in the context of an inquiry, *unfolding* of the scenario, *interpretation* of the results. Thought experimenting involves following this pattern for scenarios that are hypothetical abstract mental models, created to answer a "what if" question, and admittedly representing only a limited number of aspects of reality (see Cooper, 2005). The signaling games approach also fits this pattern: models and their assumptions represent a scenario, the consequences of the assumptions are determined by unfolding the scenario (making calculations, either analytically or running simulations) and interpreting the results.. Toy language-games and signaling game models can thus play a similar role in the CUI pattern of inquiry, despite most often being represented and unfolded using different techniques (natural language for toy language-games, and a combination of natural language, mathematics, and potentially computer simulation for signaling games). This, according to El Skaf and Imbert (2013), means that they are functionally substitutable[12].

Herein lies a possibility for additional systematicity: if signaling games can perform the same functional role as toy language-games, while enabling a more systematic approach by using mathematical language and computer simulation, one should consider them as tools to potentially supplement Wittgenstein's practice of thought experimenting. The study of language, in particular, is quite amenable to the use of mathematical models and computer simulations, and the literature is rife with them[13]. For questions of meaning, the dominant approach along these lines is formal semantics (see Portner, 2005; Winter, 2016, for introductions). But, as Stokhof (2013) points out, Wittgenstein's later philosophy is typically seen as antithetical to

[12] In particular, the authors state the following:

> We have argued that [computer simulations], [thought experiments] and [experiments], while involving different ontological types of particulars need not always serve different complementary functions; but claiming that [computer simulations], [thought experiments] and [experiments] can sometimes be functionally substitutable simply means that they can do the same thing, like two barristers or two goal-keepers are functionally substitutable but need not have the same talent. Similarly, [computer simulations], [thought experiments] or [experiments] need not, from an epistemological point of view, play their unfolding function identically nor come with warrants or credentials of the same type, have the same degree of trustworthiness or bring the same epistemological benefits. (El Skaf and Imbert, 2013, p. 3470)

[13] It is beyond the scope of this thesis to provide an overview, but Jurafsky and Martin (2009) provide a good starting point.

its core assumptions and methods:

> Arguably, formal semantics shares a number of important assumptions with views on language, meaning, and reality, and the role logic plays that Wittgenstein developed in the *Tractatus*. The distinction between the surface, grammatical form of an expression and its logical form, the all-pervading referentialism, including the defining role of truth conditions, the assumption that meaning is not only homogeneous but also universal in the sense that there can be one characterisation that applies to all (possible) languages, are some of the most important features that formal semantics shares with the Tractarian framework. This allows us to discuss the later Wittgenstein's 'criticisms on formal semantics' without actually being anachronistic. For many of the criticisms that Wittgenstein vents in the *Philosophical Investigations* against his own earlier views in the *Tractatus*, either directly or indirectly via his critique of the Augustinian picture, can be considered as criticisms of formal semantics as well, provided they are related to the assumptions that the *Tractatus* and formal semantics share. (2013, pp. 226-227)

Because of the almost lack of alternatives to formal semantics, this can lead many to believe that the *Philosophical Investigations* are antithetical with any formal approach to the study of meaning in natural language. I think that signaling games can play a similar role to Wittgenstein's toy language-games while promoting more systematicity in the study of natural language[14]. Crucially, they present an alternative to formal semantics that does not share its aforementioned problematic assumptions. As argued in the rest of this thesis, they fit well with Wittgenstein's picture of language (Chapter 4), avoid the problems of referentialism (Chapter 5), and do not conflict with Wittgenstein's remarks on rule-following (Chapter 6). They do all this while at the same time allowing for the use of mathematical modeling and computer simulations.

As El Skaf and Imbert (2013) point out, each method has its pros and cons. One advantage of mathematical models is that they make the workings of toy examples very explicit, but this sometimes can come at a cost of a loss of intuitive appeal. However, although a mathematical model allows for more control over variables and calculations, the required level of formalism can make it less flexible and thus less adequate in situations where a natural language thought experiment would suffice. Sometimes, for example when studying complex dynamical systems, mathematical

[14] When it comes to mathematical models and computer simulations in general, the argument is not completely new (see Parisi, 2004).

models supported by computer simulations are the only way to fully explore the potentially large space of possible outcomes. However, this can come at a cost of some explanatory opacity, since sometimes it is not obvious which factors contribute to which outcome (Di Paolo, Noble, and Bullock, 2000). The list goes on.[15] I believe that the advantages outweigh the limitations. Models, even if we see them as artificial and oversimplified, can serve as conversation starters. By making assumptions more explicit, they allow one to communicate their ideas more clearly to others. By exposing them, they allow their implications to be explored more thoroughly, especially when aided by unfolding mechanisms such as computer simulations.

Is the use of the signaling games framework a method for philosophical inquiry that is in line with Wittgenstein's metaphilosophical concerns? The question does not, in my opinion, have a straight answer. On the one hand, signaling game models are typically put forward to support positive hypotheses about language, which can be seen as going against Wittgenstein's recommendations, especially if one follows therapeutic or Pyrrhonian interpretations of his work. On the other hand, as I argued in Section 3.1, Wittgenstein's concerns seem to be directed more against a certain attitude towards methods than against particular methods. Positive recommendations about the right attitude to take, I argued in Section 3.2, can be drawn from making a parallel between philosophical method and explanations of meaning. With this, and particularly Wittgenstein's pluralism (Remark 3.2.3, p. 72) in mind, a further exploration of the possibility of using the signaling games framework in philosophical inquiry seems to be warranted.

Some of Wittgenstein's considerations about meaning are supported by the use of toy language-games, which are therefore implicitly endorsed. Not only that, the method is explicitly presented as a way to investigate language that helps avoiding confusion and enables a clearer picture of the issues involved (Remark 3.3.2, p. 75), with the caveat that they should not be generalized into a theory of meaning (Remark 3.3.3, p. 76). In this section, I argued that there are a lot of similarities between how toy language-games and signaling game models are used, both structurally and methodologically. This, on top of the considerations in Chapter 4, makes a strong case for seeing the use of the signaling games framework as acceptable methodology from a Wittgensteinian perspective. However, it is important to heed the warnings. One should avoid the cravings that typically lead philosophers astray and keep seeing the models as the contingent tools that they are. Although they can help explore or communicate some hypotheses about language, they should not be pursued as revealing its essence or providing universal atemporal truths about the phenomena.

[15] See Nersessian and MacLeod (2017) and Smaldino (2017) for recent overviews of the advantages and issues with the use of models and computer simulations in different areas of science.

Chapter 4

An organic picture of language

The picture of language that emerges from Wittgenstein's Philosophical Investigations *is anchored in the notion of language-games. In this chapter, I discuss this picture, how it informs a particular proposal regarding meaning and use, and whether signaling games can allow one to study language in a way that preserves these ideas.*

A core notion of the picture of language adumbrated in the *Philosophical Investigations* is that of *language-game*. There are two (closely related) ways in which Wittgenstein uses this term. Whenever clarification is warranted, I will use the term *toy* language-game to refer to the artificial scenarios of use created as a kind of thought experiment to explore particular ideas about language and meaning. Some clear examples are the so-called builder's language of §2 (extended in §8 and §86), the language of colored squares of §48 (further discussed in §64), and the well-known beetle in the box scenario of §293. I focused on this kind of language-game in Section 3.3. Another way in which Wittgenstein uses the term is to refer to the naturally occurring counterpart of these examples. They are also language-games in the sense that they involve the use of words as part of certain activities, but they are part of our existing practices rather than being artificial hypothetical scenarios. In this chapter, I focus on this notion of language-game, and on the picture of natural language that it informs.

4.1 Practice, heterogeneity, and dynamism

The term language-game is first introduced and roughly defined in §7. In the preceding sections, however, one can already find a number of toy language-games put forward to illustrate certain points. In §1, the author imagines sending someone shopping with a paper with 'five red apples' written on it, and how a shopkeeper would proceed when given such a paper. In §2, we have the often-called builder's language, where a builder and an assistant are supposed to communicate over building stones using a limited number of words; the builder calls out the words and the assistant is expected to bring him certain stones. The examples themselves involve more than just signs or sentences in the abstract. They picture situated uses of language. The discussions that follow the setup, usually revolve around the actions that the individuals involved perform (or would hypothetically perform), and around how they might have learned to respond to certain signs or stimuli in certain ways. Words are pictured as something one does things with, like tools in a toolbox (§11) or handles in the cabin of a locomotive (§12). Language itself is called an instrument (§569) that enables us to do things like "influence other human beings" and "build roads and machines, and so on." (2009, §491, p. 145e) The term language-game itself, by incorporating the word 'game' and being often compared with more standard examples of actual games (*e.g.* chess, tennis), furthermore evokes an active dimension to the concept.

Besides these implicit appeals to practice, Wittgenstein explicitly states that he is interested in looking at language in connection with practical activities, and that using language is itself part of an activity:

> We can also think of the whole process of using words in (2) as one of those games by means of which children learn their native language. I will call these games "*language-games*" and will sometimes speak of a primitive language as a language-game. (2009, §7b, p. 8e)
>
> I shall also call the whole, consisting of language and the activities into which it is woven, a "language-game". (2009, §7d, p. 8e)
>
> The word "language-*game*" is used here to emphasize the fact that the *speaking* of language is part of an activity, or of a form of life. (2009, §23b, p. 15e)

This focus on practice is important since it hints at a picture of language that is starkly different from what underlies some common philosophical reflections on meaning[1].

[1] It is impossible to do justice to all the varieties of perspectives and methodologies employed by philosophers of language when reflecting on language and meaning. The objective with this kind of loose and broad characterizations, as with the discussion that follows, is neither to create a straw man, nor to covertly refer to some author or school of thought. I want to simply highlight some hopefully recognizable features of philosophical practice that contrast with Wittgenstein's

Practices involve agents performing actions in the world. Thinking of language as a practice keeps in the forefront individuals and the potential for idiosyncratic variation; it evokes bodies, movement, and interaction; it is a first step against the urge to sublimate language onto a realm beyond the one we inhabit:

> We're talking about the spatial and temporal phenomenon of language, not about some non-spatial, atemporal non-entity. (2009, §108c, p. 52e)

Highlighting the entanglement of language with practice serves as a reminder that the activities where words and sentences are used, such as going shopping or constructing buildings, should be kept in mind when making considerations about meaning. It opposes the idea that linguistic expressions can be detached from the context of those activities, and meaning can be analyzed independently. Viewing language as a tool makes us see it as a means to an end, rather than a mere repository of information or a mirror of reality.

It also brings practical purpose to the fore: *what* we want to achieve when we use language matters to understanding *how* we use it. A typical philosophical inquiry into language and meaning would have no qualms in taking a sentence like 'Every man is mortal', 'Water boils at 100 degrees Celsius', 'Hesperus is Phosphorus', 'Snow is white', or 'The cat is on the mat', and reflect on its meaning in the abstract, not necessarily to understand that particular sentence better, but to extract general principles that govern the way language and meaning works[2]. Such an approach would ignore *who* is using that sentence, in what *context*, and for what *purpose*. Wittgenstein's objective seems to be to move away from this kind of analysis, first by situating any sort of reflection on language and meaning within the context of a practice, and second by bypassing the abstract discussion of meaning by considering

picture of language in the *Philosophical Investigations*. It is important to mention them because the author's ideas were likely developed in reaction to them, and this includes Wittgenstein's own early work.

[2] Michael Dummett gives us a clear description of this project:

> According to one well-known view, the best method of formulating the philosophical problems surrounding the concept of meaning and related notions is by asking what form that should be taken by what is called 'a theory of meaning' for any one entire language; that is, a detailed specification of the meanings of all the words and sentence-forming operations of the language, yielding a specification of the meaning of every expression and sentence of the language. (1975, p. 1)

Incidentally, one can additionally recognize here the cravings for precision (aiming for "detailed specifications"), and universality (encompassing "all the words and sentence-forming operations"), that Wittgenstein would disavow in such a project, as discussed in Chapter 3. To be fair, this is not exactly what Dummett himself defends a theory of meaning should be, but simply a description of what he calls "one well-known view". I believe Dummett's view, although more nuanced, would still be incompatible with Wittgenstein's later philosophy, but it is beyond the scope of this thesis to delve into that comparison.

instead the practical applications of words and sentences. This intention can be recognized as early as the end of §1. In the context of the 'five red apples' example, where an interlocutor asks "But what is the meaning of the word 'five'?", the narrator replies "No such thing was in question here, only how the word 'five' is used." (2009, §1d, p. 6e) Wittgenstein is not interested in questions of what words or sentences mean in the abstract, but rather in how they are used in practice. The exact nature of the relation between meaning and use will be discussed in more detail in Section 4.2. What is important to note for now is that, for Wittgenstein, in order to talk about meaning one needs to talk about use, and in order to talk about use one needs to consider language in the context of practical activities. We can summarize these ideas in the following remark:

Remark 4.1.1 *Using language is part of a practice; considerations about meaning need to take into account not only the linguistic entities but also the activities they are employed in.*

With this in mind, the unit of interest becomes not language as a whole, but language-games. Although no detailed definition of the concept is given at any point[3], §7 and §23 are certainly key. Considering what is said there, together with what one can surmise from the examples given throughout the book, it is clear that language-games have both a narrower and a wider scope than what are typically called languages. The toy language-games of §1 and §2 involve, respectively, using written signs to send someone on an errand and using speech to achieve coordination while doing construction work. A number of naturally occurring language-games are listed in §23, including giving and following orders, presenting the results of an experiment, making a joke, translating from one language to another, just to name a few. One the one hand, all of these are somewhat circumscribed activities that require knowledge of a much smaller set of linguistic elements than what makes up a whole language like English or German. On the other hand, language-games are broader in the sense of involving other things beyond a set of signs and rules for sentence formation. Buying apples or taking a bus might require a very reduced knowledge of vocabulary and grammar, but can involve other things like gesturing, making eye contact, or smiling. Although these actions would not typically be considered as part of a language, they could definitely be included as part of a language-game.

Language, in a broader sense, is portrayed as a collage or a patchwork of a large number of these language-games. A metaphor used to illustrate this at one point is

[3] And this is obviously a conscious decision (see, for example, §69) that has to do with Wittgenstein's conceptions of meaning, explanation, and his metaphilosophical position.

that of an ancient city: "a maze of little streets and squares, of old and new houses, of houses with extensions from various periods, and all this surrounded by a multitude of new suburbs with straight and regular streets and uniform houses." (2009, §18, p. 11e) One imagines language-games as different as neighborhoods from various periods in a city: some more organically developed, others following a strict high-level plan; some more baroque with many purely decorative elements, and some more minimal and functional in spirit. Despite the variability, language-games are nevertheless linked together in making up what one calls a language, the same way that different boroughs make up a city. The objective of this metaphor is to highlight the heterogeneity of language in terms of the diversity of language-games that compose it.

The analogy also brings out another important aspect of Wittgenstein's picture of language: dynamism. Observing how different parts of the city developed in different periods is not only a historical curiosity, it is also a reminder that it will most likely keep developing in various ways in the future. As with the ancient city, so too happens with language:

> But how many kinds of sentence are there? Say assertion, question and command? – There are *countless* kinds; countless different kinds of use of all the things we call "signs", "words", "sentences". And this diversity is not something fixed, given once for all; but new types of language, new language-games, as we may say, come into existence, and others become obsolete and get forgotten. (2009, §23a, pp. 14e-15e)

We should never assume that a natural language is complete, but rather expect it to keep changing over time, with new language-games being added, and old ones falling out of use. The number of language-games that compose a language is thus, for all practical purposes, countless, not only for their sheer number and variety, but for their dynamic nature as well. The following remark summarizes the ideas discussed so far:

Remark 4.1.2 *Language is deeply heterogeneous and dynamic, consisting of an ever-changing multitude of language-games.*

The metaphors just discussed, although highlighting heterogeneity and dynamism, can be somewhat misleading in another respect. One could get the idea of language-games as clearly bounded and separate from each other, only connected at the edges like the patches of a quilt or the boroughs of a city. This is not, however, how Wittgenstein envisions the relation between language-games, and this is quite patent in the discussion starting in §65. Language-games are said to have all sorts of affinities between them (§65), characterized as a "complicated network of similarities overlapping and criss-crossing: similarities in the large and in the small." (2009, §66, p. 36e) Thus, two language-games can be similar to each other but different from a third with respect to one aspect, but this third can be similar to the first and dissimilar to the second according to another aspect. The relation between language-games is thus seen as multi-dimensional, breaking with the two-dimensional metaphors of the patchwork and the city.

It is important to notice that the idea of multi-dimensional overlapping and criss-crossing similarities raises the possibility of some language-games being hard to distinguish from others. This can happen, for example, when the similarities are many and the differences depend on subtleties, like tone of voice or facial expressions (see §21). We can all recognize this possibility by considering examples like when an ironic remark is taken literally, or when a joke goes over someone's head: one interlocutor might have been trying to play one language-game which the other failed to identify. Because of this, and unlike what the metaphors of a patchwork or a city could lead us to believe, boundaries between language-games can be hard to draw.

Wittgenstein is very adamant about the idea that there is not, however, necessarily anything in common between *all* language-games. This is the main point of §§65-67, where the examples of 'game' and 'number' serve to illustrate the idea that one can use the same word to talk about things that do not all necessarily share a common feature. Things in such a situation are said to have a *family resemblance* between them and thus form a family. The analogy is with how people who are blood relatives can share similarities and also have differences, depending on which aspect, like eye color, type of hair, nose shape, and so forth, one is considering, without necessarily there being one characteristic that is common to all. One could nevertheless talk about them as a family, just like one can talk about a language as composed of many interrelated language-games without there being a common aspect that all those language-games share. We can summarize these points in the following remark:

Remark 4.1.3 *Language-games are interrelated along many different dimensions, without anything necessarily common to all.*

The remarks summarized in this section paint a picture of language as organic: heterogeneous, ever-changing, adaptable. I use the term 'organic' in a loose sense. I want to suggest that language, in this picture, shares similarities with the results of living processes. For example, the process of evolution by natural selection, as we understand it today, has created a heterogeneous set of interrelated species that continuously change and adapt to their environment. One can think of language in similar terms.[4] Wittgenstein's notion of language-game is a key component of this picture that additionally evokes the interconnection between language and practice. Words are portrayed as instruments used to enable the achievement of certain goals. In the context of this picture, there is little room for a notion of meanings as entities standing in a relation of correspondence to linguistic expressions. I discuss the relevant remarks for this negative claim in more detail in Chapter 5. In the following section, I want to first explore remarks Wittgenstein makes about what seems as a positive proposal of a way of thinking about meaning that goes along with this organic picture of language.

4.2 Meaning and use

Wittgenstein's general pragmatist stance when it comes to language feeds into a particular conception of meaning in relation to use. Though no sequence of sections in the *Philosophical Investigations* is dedicated to the topic, one observation that repeatedly appears in the discussion of other topics is the importance of understanding use in order to understand meaning. This observation occurs often when Wittgenstein talks about learning a language or a language-game. For example, in §§5-9 one is lead to imagine what would be involved in the teaching of the builder's language of §2 (and an extension thereof laid out in §8). The exercise serves to expose the limitations of ostensive teaching. In drawing attention to those limitations, Wittgenstein suggests that learning a language is not a matter of explaining but training (§5b). An important part of this training might involve the teacher pointing to an object and uttering a word (§6). This training could be interpreted as having the effect of establishing a connection between the object and the word. However, Wittgenstein counters the following:

> But if this is the effect of the ostensive teaching, am I to say that it effects an understanding of the word? Doesn't someone who acts on the call "Slab!" in such-and-such a way understand it? – No doubt it was the ostensive teaching that helped to bring this about; but only together with a particular kind of instruction. With different instruction the same

[4] This has been more recently formulated, for example by Beckner et al. (2009), in terms of the notion of a complex adaptive system. I briefly discuss this idea again in the conclusions.

> ostensive teaching of these words would have effected a quite different understanding. (2009, §6c, pp. 7e-8e)

What is being drawn attention to is that, although instilling an association between object and word in the learner might be important in teaching the meaning of the call "Slab!", it is not enough. The purpose of the call in the language-game is not to evoke mental imagery, but to elicit a certain behavior (§6b). If someone only learned the connection between the object and the word, this would not be enough to know how to get someone to pass him a slab, or know what to do when hearing the call "Slab!".

This point is illustrated again in §31, with an example from chess: we can only be said to understand what the king is, not by naming it, but by knowing what to do with it, how to move it around the board, and so on. In §§40-44, the case is made that for a proper name to be said to have a meaning it is quite enough that it has a use in the language-game, even if the bearer of the name, for example, ceases to exist. The discussion in §§139-142 draws attention to how a picture of a cube is, in most cases, insufficient for one to be able to use the word 'cube' successfully. In the 'beetle in the box' though experiment[5] (§293), what seems to be crucial about the word 'beetle' having meaning is not the object in each person's box, but that the word has a use nonetheless. The idea being called into question throughout these remarks is the idea that "[o]nce you know *what* the word signifies, you understand it, you know its whole application" (2009, §264, p. 100e). This is repeatedly challenged throughout the book, sometimes more explicitly as in the examples just discussed, sometimes more implicitly. Whenever Wittgenstein seems to be making a point about what meaning is not, reminders about use are always present.

In addition to the importance of use in language acquisition, Wittgenstein also makes remarks regarding how the meaning acquired through such a process is maintained. To this respect, we are often reminded that signs, words, or sentences do not *have* meaning in and of themselves. This idea, although tempting, is quickly dismissed in passages such as the following:

> I am told: "You understand this expression, don't you? Well then – I'm using it with the meaning you're familiar with." As if the meaning were an aura the word brings along with it and retains in every kind of use. (2009, §117a, p. 53e)
>
> Every sign *by itself* seems dead. *What* gives it life? – In use it *lives.* (2009, §432, p. 135e)

[5] The experiment is discussed in more detail in Section 5.3.

> How does it come about that this arrow ↣[6] *points*? Doesn't it seem to carry within it something extraneous to itself? – "No, not the dead line on paper; only a mental thing, the meaning, can do that." – That is both true and false. The arrow points only in the application that a living creature makes of it. (2009, §454, p. 140)

Vibrations of air molecules or ink marks on paper (the material things we exchange when using language) do not carry with them or within them something extra that endows them with meaning. The dismissal of these metaphors in these passages is accompanied by appeals to both use and to the organic picture of language discussed in the previous section. The objection in the first quote from §117a is even more significant if we consider the vision of language as diverse and dynamic.

These aspects of language make it even harder to believe that an expression could contain in itself, not only the meaning for every known use of the word, but also potentially all future uses that could ever be invented. The second quote from §432 suggests that signs come to life during use. The image of life appears again in different form in the third quote from §454, where it is said that a *living* creature is necessary to make an arrow point. These associations reinforce the idea of language and meaning as something dynamic. The third quote again evokes the picture of words and language as tools, something a creature makes use of, and is thus a reminder that language is part of a practice. Meaning does not exist outside of signs and expressions constituting with them a special bond, nor is it a static property or comes attached to them in some mysterious way. An arrow points when it is used to point. Note that these are again grammatical remarks about meaning and relate to the way that we talk about signs, rather than to a hypothetical metaphysical entity. Thus, it is not that a sign has meaning when is it used in a certain way, but that we say of a sign that it has meaning when is it used in a certain way. These points can be summarized in the following remark:

Remark 4.2.1 *Signs do not carry or possess meanings; we say that they have meaning when they are being used.*

Clear statements regarding the exact nature of the relation between meaning and use are rare. The well-known passage that does appear to provide this reads as follows in the revised translation by Hacker and Schulte:

> For a *large* class of cases of the employment of the word "meaning" – though not for *all* – this word can be explained in this way: the meaning of a word is its use in the language. (2009, §43a, p. 25e)

[6] This is not an accurate rendering of the arrow portrayed in the book.

The interpretation of this remark is, however, slightly contentious. One can distinguish between two main interpretations. The first takes the passage as defending the idea that we can *define* meaning as use. A strong advocate of this proposal is Paul Horwich (*e.g.* 2004; 2008) who defends what he calls a Use Theory of Meaning (UTM). In his reading, §43a is taken as clear evidence that Wittgenstein is concerned with "the facts in virtue of which a given word has the meaning it does—with the underlying characteristics that are responsible for its possessing that particular meaning" (2008, p. 134). The suggestion is thus that words have particular meanings and that these are grounded in facts related to their use. This, according to UTM, applies in general, *i.e.* "the meaning of a word, *every* word, is its use" (2008, p. 138). Under this interpretation, a research program that could "discover the particular meaning-constituting use-properties of particular words" (2008, p. 137, footnote 3) is a plausible endeavor, albeit one for linguistics, not philosophy.

This interpretation seems to me to be at odds with many aspects of Wittgenstein's later philosophy. First, it aims to be a general characterization of the essence of meaning. This goes against the anti-essentialist picture of language as a heterogeneous patch of language-games linked by family resemblances (Remark 4.1.3, p. 94). To declare a definition of meaning as use is to fail to see Wittgenstein's words in §43a in the broader context of the conception of language that runs through the rest of the book. It is also to fall into the temptation of trying to "*see right into* phenomena" (2009, §90, p. 47e) and thinking there is "something that lies *beneath* the surface" (2009, §92, p. 48e), a mistake Wittgenstein repeatedly warns us not to make. Second, UTM construes words as having definite meanings, something that goes against Wittgenstein's rejection of the idea of meaning as something that accompanies, or that is related to, words and linguistic expressions (Remark 4.2.1, p. 97). Talk about words "possessing meanings" or having "meaning-constituting use-properties" is, it seems to me, to go against those remarks. Third, such a strong conception of the connection between use and meaning readily invites the idea that the former can determine the latter. This is something that Wittgenstein himself explicitly questions at time, for example in §§139-141 using the example of the word 'cube'[7], and more generally in §§191-197.

For these reasons, I think the formulation "the meaning of a word is its use in the language" in §43a is somewhat misleading. When taken out of context, it could be interpreted in a way that runs counter to a number of aspects of Wittgenstein's later philosophy. But what is being suggested in §43 is not a theory of meaning. It is simply a reminder that the meaning of a word is usually explained by characterizing

[7] For more on this example, see discussions leading up to Remark 6.1.2 (p. 149) in Chapter 6 of this thesis.

its use in the language, just like "the *meaning* of a name is sometimes explained by pointing to its *bearer*." (2009, §43b, p. 25e) This does not imply that the meaning of that name actually *is* the bearer, neither should we in general identify the meaning of a word with its use. Martin Stokhof (2013) also rejects UTM as something that Wittgenstein would not have endorsed, and does so on similar grounds. He recognizes a different possible reading of §43a, namely as more of a methodological advice:

> The minimalistic interpretation of 'explaining meaning by looking at the use' reads it as a purely methodological statement. On this view it is not so much a connection between meaning and use that is made, but a shift of attention that is effected. It invites us to stop looking for some 'thing' that we can call meaning, and focus instead on the way expressions are used: that should suffice. (2013, p. 223)

The interpretation is that the point of §43a is not to identify meaning with use, but to suggest that one should divert questions about meaning to discussions about use. Rather than falling prey to the temptation of trying to answer questions of the form "What is the meaning of X?" as "The meaning of X is Y", one should instead investigate the ways in which 'X' is used. And this is not meant as a first step towards later identifying those as the meaning of 'X', but simply as a different method of answering the question. For example, when facing the question "What is the meaning of the word 'game'?", how should one proceed? One could feel tempted to provide a definition; a philosopher would most likely try to investigate the necessary and sufficient conditions for something to be called a 'game'. Wittgenstein's advice goes in a different direction:

> How would we explain to someone what a game is? I think that we'd describe *games* to him, and we might add to the description: "This *and similar things* are called 'games'." (2009, §69, p. 37e)

Describing activities that one would call 'a game' is a way of describing the use of the word 'game'. It is also, for Wittgenstein, a perfectly acceptable way to explain the meaning of the word 'game', and thus to try to answer the question "What is a game?" without providing any type of precise definition.

This methodological advice is similarly recognized by other authors. For example, by Richard Rorty when he defends that the association of meaning and use proposed in the *Philosophical Investigations* "is not a 'use-theory of meaning', but rather a repudiation of the idea that we need a way of determining meanings." (2007,

p. 172) This methodological interpretation strikes me as the best fit with Wittgenstein's later philosophy: it rejects the confusion of thinking that meaning must be a property (or a set therefore) of linguistic expressions; it fully recognizes the heterogeneous and dynamic nature of language; and it acknowledges the importance of use without constraining it into an essentialist characterization of meaning.

Remark 4.2.2 *Meaning is not constituted by use, but in order to understand meaning we should study how language is used.*

Wittgenstein's remarks about meaning and use, when taken together with the rest of his later philosophy, should not be interpreted as endorsing any sort of identification between the two: it is not that meaning *is* use, but rather that we should always reflect upon meaning *in* use. Learning or teaching the meaning of words or other linguistic expressions always involves learning or teaching some use for them; language can only said to have meaning when it is being put to some use. Thinking of meaning in these terms aligns with Wittgenstein's appeals to the primacy of practice and fits well with the organic picture of language advocated throughout the book.

4.3 Signaling games and the organic picture

In order to ascertain whether or not, and to what extent, the framework of signaling games fits with Wittgenstein's later philosophy of language, it is crucial to consider how the approach fares in light of the remarks discussed so far. The study of signaling game models aims to illuminate hypotheses about certain features of natural language. The framework provides guidelines to construct abstract models that are more akin to Wittgenstein's toy language-games than to their natural counterparts (see Chapter 3). However, in providing abstractions, they purport to capture some features of real-life practical activities where use of language is involved. As such, the assumptions embodied in their implementation, and the methodology used to study them, can be subject to comparison with Wittgenstein's remarks.

One key focus lies in the characterization of language in terms of practice (Remark 4.1.1, p. 92). As discussed above, this serves as a reminder that linguistic entities do not stand as an object of study on their own. They are tools used by individuals situated in a context and driven by a purpose. These elements are strongly intertwined, and therefore should not be abstracted away. Practices are multifarious and complex. Signaling games embody these lessons very well. When creating a model according to the framework, the researcher includes and characterizes senders and receivers (or populations thereof), a context (states, messages, actions, and possibly other elements), and a purpose (captured by the utility function). These elements form a system of interlocking parts, and the questions that drive the study of the models almost always revolve around how (and to what extent) agents, context, and purpose interact and influence signal use. Section 2.2 gives an impression of the extent to which this motivates a variety of explorations that go beyond simple considerations of linguistic entities and their meanings. This includes studying various types of agents and how they are connected, different number and distributions of states, signals, and actions, dynamic versus static contexts, aligned versus conflicting purposes, and many others. The variety of issues explored by researchers working with signaling games comes from taking the phenomenon of meaning as an irreducible interplay between all the elements that constitute language as a practice. The ability of the framework to provide insight into them attests to the benefits of taking this picture seriously.

Much like with Wittgenstein's language-games, each individual model captures something, on the one hand, broader, and on the other hand, narrower than what we typically call a language. Broader because it includes more elements than the linguistic entities, narrower because the scope of each model is more limited. One author can explore a model where agents exchange a signal in order to coordinate on a place to meet, and another can devise a signaling game where agents have a conflict of interest. There are models where states are distributed uniformly, and there others designed to study the impact of skewed priors. That these example scenarios are difficult, or potentially impossible, to reconcile under a more general model is, however, not usually seen as a problem. And it would only be so if one was looking for something like a general theory of meaning. On the methodological level, researchers working with the signaling games framework can be seen as embracing the heterogeneity of language (Remark 4.1.2, p. 93), first by focusing on particular cases of use, and second by not attempting to coerce all possible scenarios into the same model. They can be seen as appreciating the perspective that these models, like different language-games, can be related but need not have something common to them all (Remark 4.1.3, p. 94).

When it comes to heterogeneity, one should also reflect on how it is handled on a more technical level. Most models in the literature represent one specific game. Agents, or populations thereof, engage in interactions characteristic of a well-defined situation. Not everything is, however, set in stone. One source that creates some room for variation is parameterization. Most models have characteristics that can be changed by setting different values to a parameter. Drawing again on the models discussed in Section 2.2, think of something as simple as the number of states in the state space, or other aspects like the bias in a prior distribution, the cost of sending a given message, the perceptual acuity of agents, the average number of connections in the social network of the population, and so forth. The possibilities of what to parameterize in a model are typically only restrained by the imagination of the researcher. Each value of a parameter specifies an instance of the model that can be seen to characterize a slightly different scenario of signal use. By stemming from the same model, these different instances are nevertheless strongly related to each other along various dimensions. This is one additional way in which signaling games can be seen as promoting the embrace of heterogeneity and Remark 4.1.3 (p. 94).

If a given feature of the model is chosen to be left as a parameter, it is important to study the effect of that parameter in the outcome of the analysis of the model. This is typically done by studying model instances independently from each other and subsequently comparing the results in terms of external metrics. With this approach, however, heterogeneity is not fully being taken into account. Each instance of the model is studied in isolation from the other. However, it is important to realize that, in natural language, all these variations can simultaneously coexist and this can have an impact on signal use. Even in the context of a single model, studying a truly heterogeneous environment can reveal unexpected effects.

An example of this can be found in the literature on vagueness. Both O'Connor (2014b) and Franke and Correia (2018) propose signaling game models where vague signal use is shown to develop as a result of, respectively, a type of learning dynamic and an imitation-based evolutionary process. Each model has its own parameter conditioning the degree of vagueness observed in the states of equilibria. Various parameter values are analyzed for each adaptive process. In both cases, based on comparing independent runs in terms of external metrics, the authors hypothesize that certain values of the parameter that induces a degree of vagueness stimulate faster convergence on a coordinated signal use. This, in turn, is hypothesized to potentially allow agents that tend to evolve a communication system with some vagueness to dominate over agents that use those signals more precisely, because they can temporarily attain an advantage. However, as Correia and Franke (2019) show[8], when one studies one of these models in a more heterogeneous environment—using a multi-population variant of the replicator dynamic—unexpected interactions between populations with different parameter values can make this story much more complicated than it may seem from studying homogeneous environments indepen-

[8] See Appendix A.

dently.

There is another interesting way in which this relation between language-games can be taken into account in signaling game modeling. When thinking of language-games played by rational agents, one needs to consider the level of awareness the agents have of the situation they are involved in. As mentioned in relation to Remark 4.1.3 (p. 94), given the potential close similarities between slightly different language-games, it is possible for one agent to believe he is playing one game, while the other agent thinks he is playing another. This kind of scenario is studied in the game theory literature as hypergames, a notion first introduced by Bennett (1977). Some recent work has explored evolutionary dynamics for these models (Kanazawa, Ushio, and Yamasaki, 2007; Jiang et al., 2018), as well as their relationship with Bayesian games (Sasaki and Kijima, 2012, 2016). To the best of my knowledge, no direct applications of this literature to signaling games have been studied yet. However, there is at least one independent proposal of so-called games with unawareness that serve the same purpose: they represent situations where agents might not be fully aware of the game they are playing, and consider the implications that the same happens to the agent they are playing with. Franke (2014b) advances these models as ways of studying pragmatic inferences of the Gricean kind. The signaling games framework thus has the tools to represent and study such complex relations between language-games as those foreseen by Wittgenstein.

Remark 4.1.2 (p. 93) mentions not only heterogeneity but also dynamism. Going back to the methodological level, recognition of the dynamic aspect of language is usually present in the signaling game approach, although this is intimately tied to the way models are analyzed. Lewis approached his models along the lines of classic game theory, *i.e.* by providing a story about which strategy or strategies a rational player would choose in the situation characterized by the game. Note that in this type of analysis there is no room for considerations of dynamism in signal use, because the models are seen as single shot (or single stage) games. The choices of which signal to use (by the sender) and of which action to take (by the receiver) are one-off, and do not have any temporal dimension. The processes of rational deliberation that lead up to the choices can be iterative, but the actual signal exchange is seen as a single event, potentially independent from other instances. Single shot games are typically analyzed using static notions like the Nash equilibrium. Interestingly, the notion of an evolutionarily stable strategy, although motivated in dynamic terms, is actually also static, because the story about potential invasions by other strategies is simply considered in the hypothetical.[9]

[9] For more on these notions, see again Section 2.1.

In order to take the dynamic aspect of language into account, one needs to study repeated games, *i.e.* situations where a single shot game is played several times by either the same agents or the same populations of agents. This gives the opportunity to consider strategies that can be adapted from one instance to another based on previous successes or failures. As discussed in Chapter 2, adaptive processes can be motivated in terms of evolution (biological or cultural), learning, or rational deliberation. The important thing to note here is that, because of the repeated nature of the interaction, signal use can change over time. This can happen even in a static environment, simply as sender and receiver strategies mutually adapt to each other. Static notions like the Nash equilibrium can inform the analysis of such models, for example by indicating some attractors of the system. But ultimately these games should also be studied using dynamic equations or agent-based simulations, approaches which can reveal much more about the adaptive processes and how they drive the system towards equilibria, if at all[10]. Some systems, like the simple binary signaling game, can almost always[11] be driven to one of the two signaling systems, but others can have local attractors that pull strategies away from reaching global optima, or forever be stuck in cycles, like in the signaling variant of rock-paper-scissors (*e.g.* Wagner, 2012).

But taking the dynamic nature of language seriously means going even further beyond the notions of attractors and equilibria. Wittgenstein sees language as continuously changing, rather than evolving towards some eventually stationary state and stopping there. Awareness of this possibility creates additional motivation to understand how information exchange can still occur outside of stable equilibria, as in the work of Wagner (2012), to consider scenarios where the environment itself is regularly changing, as in the work of Alexander (2014), and to study how change sparked from creativity propagates in a network, as in the work of Mühlenbernd (2017). These examples show how the signaling games framework lends itself to deeply embracing the dynamism of communication, and how that helps illuminating some of the questions arising from that perspective.

The majority of questions addressed using the signaling games framework, when translated in the technical terms of the model, take the form of considerations about strategies. The researcher is typically interested in knowing, given a characterization of a certain activity as a concrete model in terms of agents, context, and utility, how different strategies fare against each other, and which ones are favored by some adaptive process or another. Strategies are representations of either which signals

[10] For a more detailed argument on the importance of analyzing signaling game models using dynamic approaches, see Huttegger and Zollman (2013).

[11] In a mathematical sense, *i.e.* it happens with probability 1.

are produced by an agent (or a population thereof) given a certain state of affairs, or which actions are undertaken upon exposure to a certain signal. They are, therefore, specifications of signal *use*. The important thing to notice here is that this is usually the focus of analysis within the signaling games framework. Here is how Lewis talks about meaning after characterizing the Paul Revere signaling game[12]:

> I have now described the character of a case of signaling without mentioning the meaning of the signals: that two lanterns meant that the redcoats were coming by sea, or whatever. But nothing important seems to have been let unsaid, so what has been said must somehow imply that the signals have their meanings. (1969, pp. 124-125)

This passage can be seen[13] as expressing a similar idea as Remark 4.2.2 (p. 100), *i.e.* that in order to understand meaning one need not conceive it as an entity, but can rather focus on studying language use. Signaling games provide a way to do so. Rather than focusing on what signals mean, researchers direct their attention to how they are used. We are still talking about meaning, but without having to hypostatize meanings. I will return to this point in Chapter 5.

Remark 4.2.1 (p. 97) sheds additional light on the matter. As mentioned there, linguistic entities do not have meaning in and of themselves. In the signaling games framework, messages do not have any intrinsic meaning either. What intuitively seems, from a third person perspective, as a meaningful use of signals depends first on the strategies, *i.e.* on how signals are being made use of by the agents. Consider the simple binary signaling game of Section 2.1 and the example strategy profiles represented in Figure 2.1. Signaling systems are combinations of strategies which one would intuitively describe as meaningful: whenever a certain state obtains the sender chooses one and the same message given which the receiver chooses the optimal action, and conversely for the other state. Perfect coordination between state and action is achieved via univocal exchange of messages. However, pairing the sender strategy of one possible signaling system with the receiver strategy of the other leads to complete miscommunication (anti-signaling). Having separate strategies for sender and receiver already allows one to notice something important: what looks like a meaningful message exchange depends on how *both* parties behave.

If one further considers the dynamic aspect of language and how it is taken into account in the signaling games framework, another dimension of meaning is

[12] An example of a signaling game based on the famous code established by Paul Revere and the sexton of the Old North Church in Boston to signal the movements of English troops before the start of the American Revolutionary War (see Lewis, 1969, pp. 122-125).

[13] Although other passages in the book are somewhat at odds with this interpretation. More on this in Section 5.4.

$$\begin{tikzcd}
t_1 \arrow[r, ".21"] \arrow[dr] & m_a \arrow[r, ".46"] \arrow[dr] & a_1 \\
t_2 \arrow[r, ".63"'] \arrow[ur] & m_b \arrow[r, ".11"'] \arrow[ur] & a_2
\end{tikzcd}$$

Figure 4.1: Example of a random mixed strategy profile for a simple binary signaling game. Values can be read as probabilities and add up to 1 per choice point. In this example, $\sigma(t_1, m_a) = .21$ indicates that the probability of m_a being chosen by the sender on state s_1 is .21, and this implies $\sigma(t_1, m_b) = .79$.

revealed. When a signaling game is analyzed in terms of an adaptive process, one usually starts with random strategies for both sender and receiver. An example is the mixed strategy profile represented in Figure 4.1. In such a profile, messages and actions are chosen with a certain probability, but without any clear systematicity; one would hardly call such an exchange meaningful. As discussed in Section 2.1, adaptive processes like reinforcement learning or the replicator dynamics can be shown to gradually drive such strategy profiles into a signaling system. However, there is no point in such a gradual process where signals are imbued with meaning, they are contentless throughout. Yet, the system goes from a state where there is no apparent meaningful exchange of messages, like what is represented by the strategy profile in Figure 4.1, to a signaling system, where systematic coordination through the exchange of messages is achieved by the agents. This supports the idea that meaning is not a property of a linguistic entity, but that meaning arises *in use.* And it further strengthens the observation that, in order to understand meaning, one needs to study a whole system of agents, context, purpose, and dynamic adaptation.

In conclusion, the *Philosophical Investigations* contains a number of remarks that can be seen as constituting a picture of language as organic: heterogeneous, dynamic, and adaptive. This is partially anchored in the notion of language-games as practices that involve the exchange of linguistic entities by agents situated in a context and driven by a purpose. What we call natural language can be seen as a family of an ever-changing myriad of such language-games, related to each other along several dimensions, but without anything necessarily common to all. Within this picture, meaning is related to, but not constituted by, use. To better understand meaning, one should direct our attention to studying use. The signaling games framework fits well with this perspective. If we see each signaling game model as attempting to capture a particular language-game, heterogeneity and dynamism are usually well embraced on a methodological level

On a technical level, this depends on the type of analysis conducted, which varies between individual approaches. More can be done to explore the interrelations between similar language-games and the truly continuous dynamics of natural language, but some work in the literature shows that both technical tools and interest in the topic exist. With regards to the shift of perspective from seeing meaning as entities to studying use, the framework seems to be much more strongly aligned with the latter. It is important to note that there are many aspects of natural language that have not yet been considered in the signaling games literature. This is to be expected in a somewhat young research approach dealing with a highly complex phenomenon. Embracing the organic picture of language that Wittgenstein presents seems to me as the best way to handle this complexity. It motivates an approach that is piecemeal, in that it directs its attention to narrower aspects of the whole phenomenon at a time, but also open, in that does not attempt to make hasty generalizations or universal claims. But, in order to defend these observations, one should further reflect on the many remarks concerning method in *Philosophical Investigations*. This is the focus of the next chapter.

Chapter 5

It's not the word that counts

The Philosophical Investigations *contains a criticism of the notion of meaning as anchored in correspondence. In this chapter, I explore Wittgenstein's remarks against different varieties of this idea, and how to keep those concerns in mind when interpreting signaling game models.*

A common, even perhaps intuitive, idea about meaning is to conceive of it in terms of a relation between linguistic entities (words, expressions, sentences) and other things. If we can ask "what is the meaning of so-and-so?" it seems somewhat natural to expect the possibility of a reply of the form "the meaning of so-and-so is this-or-that". This way of talking about meaning motivates the idea that the meaning of a linguistic entity is some other entity that is somehow related to it. This is what Wittgenstein describes, somewhat mockingly, in the following passage:

> People say: it's not the word that counts, but its meaning, thinking of the meaning as a thing of the same kind as the word, even though different from the word. Here the word, there the meaning. The money, and the cow one can buy with it. (2009, §120f, p. 54e)

This is possibly an intuition that guides everyday considerations about meaning. But the reason Wittgenstein is interested in this picture is most likely because it also underlies many philosophical discussions. Especially since the late 19th century[1], philosophers have tried to make their assumptions about language more explicit by developing theories of meaning. Philosophical discussions about meaning usually revolve around the nature of the entities to which expressions of language correspond. But the idea that for an expression in a language to have meaning is for it to correspond to something else has largely gone unquestioned throughout.

[1] See Rorty (1967).

Even nowadays, one can see how this picture is ingrained. Speaks, for example, identifies two sorts of theories of meaning:

> The first sort of theory—a semantic theory—is a theory which assigns semantic contents to expressions of a language. [. . .]
>
> The second sort of theory—a foundational theory of meaning—is a theory which states the facts in virtue of which expressions have the semantic contents that they have. (2018)

The task of the former is to determine what meaningful linguistic entities correspond to, whereas the former investigates how such a relation was established, but both types of theories presuppose that for a linguistic entity to have meaning is for it to have semantic content. I have briefly discussed how Wittgenstein rejects this picture of meaning in general (see especially the arguments leading up to Remark 4.2.1, p. 97). In this chapter I want to explore Wittgenstein's objections against particular incarnations of this idea. This includes three varieties: meanings as everyday objects (Section 5.1), meanings as metaphysical simples (Section 5.2), and finally meanings as mental entities (Section 5.3). Wittgenstein does not strictly separate the criticism of these three varieties. Furthermore, the target is usually the general picture of meaning as correspondence. Because of this, even though I separate the discussion that follows in three sections, remarks are not exclusively relevant to only one variety and can often be applied to others.

It should be noted that theories of meaning have grown more elaborate since Wittgenstein wrote the *Philosophical Investigations*. It can therefore seem that the three varieties of the picture of meaning as correspondence are not relevant any more. As mentioned before, I believe that Wittgenstein criticism is aimed at the general conception of meaning as correspondence and, insofar as this picture still underlies many contemporary theories of meaning, his remarks probably still apply. However, it is not the purpose of this chapter to make that argument. I look into the criticism in more detail in order to understand how they may point beyond them, but I only focus on they apply to the signaling games framework (Section 5.4).

5.1 Everyday objects

The *Philosophical Investigations* opens with a quote from Augustine describing how he recalls learning language as a child. This serves to illustrate an idea about how language works that Wittgenstein summarizes as follows:

> These words, it seems to me, give us a particular picture of the essence of human language. It is this: the words in language name objects – sentences are combinations of such names. — In this picture of language we find the roots of the following idea: Every word has a meaning. This meaning is correlated with the word. It is the object for which the word stands. (2009, §1b, p. 5e)

This is a clear characterization of a possible variety of the picture of meaning as correspondence: the idea that the meaning of words like 'table', 'chair', and 'bread' are just those everyday objects that these words refer to. I will henceforth call this position *externalism*[2]. The rest of §1 follows with a number of remarks that prefigure the discussion up to §38. The first (§1c) is the reminder that this idea is probably inspired by considering only a certain kind of word, namely common and proper nouns, but that other kinds of words also exist and the picture portrayed might not work well for all of them. In the remaining paragraph (§1d), a first instance of what could be called a toy language-game (see Section 3.3) is introduced. The imagined scenario is one where someone is sent shopping with a slip of paper that has "five red apples" written on it, followed by a possible account of how the shopkeeper could operate with those words. In this interchange, one can find hints of the question of how one knows what to do with words, and the suggestion that understanding the use of a word is enough to understand its meaning (on this, see again Section 4.2).

The first way Wittgenstein resists this picture of meaning is to remind the reader of the heterogeneity of language (see again Section 4.1). These reminders are not necessarily a full-blown refutation, but they certainly contribute to making us see the limitations of the idea. They are, at the very least, an attack on externalism as a *general* theory of meaning. This is quite clear in §§2-5, for example in the following passage:

> Augustine, we might say, does describe a system of communication; only not everything that we call language is this system. And one has to say this in several cases where the question arises "Will that description do or not?" The answer is: "Yes, it will, but only for this narrowly circumscribed area, not for the whole of what you were purporting to describe." (2009, §3a, p. 6e)

[2] In philosophy of language, this label is typically attributed to a particular theory of meaning, made famous by Kripke (1972) and Putnam (1975) (see Hale and Wright, 1997, p. 681), restricted to proper names and indexicals. It would be anachronistic to say that Wittgenstein is addressing this theory. However, the underlying intuition behind it is very similar to the one stated here, thus I think it is useful to slightly abuse the label here.

Here it is granted that externalism might work for a certain type of use of words, but potentially only for that particular case, not for the whole of language. This type of reminder, and an initial description of the risk of ignoring it, show up again in §§11-14. In these remarks, Wittgenstein compares words with tools in a toolbox and handles in the cabin of a locomotive. Like with those objects, the spectrum of types of use of different words is broad. However, because words appear similar to each other (when written in print, or in speech), one can fail to see that. The risk is then to take one type of use, like making reference to an object, and assume the rest of language functions in the same manner. One would either get an account of meaning that doesn't work for most linguistic expressions, or, in order to try to make it fit, create generalizations so broad that they become useless (§§13-14).

Not only is there a variety of uses of signs, words, and sentences, but also that their applications are countless and language is ever-changing, with new language-games being created and others being forgotten (§23). In §24 there are again hints at the repercussions that a single-minded view can have in the context of this heterogeneity: if one considers externalism as a general theory of meaning, one will be prone to apply its model to language-games where it doesn't fit, and potentially be lead to create pseudo-problems that would not otherwise arise. In §§26-27 it is suggested again that naming is not enough to characterize everything that we do with words, with examples of one-word exclamations used to illustrate the point. The reminder is always the same: language is more heterogeneous than the cases externalism could apply to, thus a one-size-fits-all approach might not do. This leads us to our first remark of this section:

Remark 5.1.1 *There are more kinds of words than nouns, more kinds of sentences than assertions, more ways to use language than naming objects.*

In order to further chip away at the assumptions behind externalism, Wittgenstein introduces a toy language-game where this picture of meaning apparently applies:

> Let us imagine a language for which the description given by Augustine is right: the language is meant to serve for communication between a builder A and an assistant B. A is building with building stones: there are blocks, pillars, slabs, and beams. B has to pass him the stones and to do so in the order in which A needs them. For this purpose they make use of a language consisting of the words "block", "pillar", "slab", "beam". A calls them out; B brings the stone which he has learnt to

> bring at such-and-such a call. — Conceive of this as a complete primitive language. (2009, §2b, p. 6e)

The scenario just described is one where it would seem natural to say that the meaning of 'block' is a particular building stone, and that the word refers to those objects. Some of the remarks that follow reflect on how such a language could be taught to a child. One possible method would be simple ostensive teaching, which, in the case of this particular language-game, would consist in "the teacher's pointing to the objects, directing the child's attention to them, and at the same time uttering a word" (2009, §6b, p. 7e).

Even if we concede that this method could establish a relation between words and objects for a particular child, by itself it would not be enough to teach the child to use the language successfully:

> This ostensive teaching of words can be said to establish an associative connection between word and thing. But what does this mean? Well, it may mean various things; but one very likely thinks first of all that a picture of the object comes before the child's mind when it hears the word. But now, if this does happen – is it the purpose of the word? – Yes, it *may* be the purpose. – I can imagine such a use of words (of sequences of sounds). [. . .] But in the language of §2 it is *not* the purpose of the words to evoke images. (It may, of course, be discovered that it helps to attain the actual purpose.) (2009, §6b, p. 7e)

Imagine a child that has learned the language in this way and interacts with a builder. She hears the word 'slab' and imagines an object but does nothing. Would the builder think that she understands the word? Would we like to say that she knows the meaning of 'slab' if she does not bring a slab to the builder? Wittgenstein does not here deny that the ostensive teaching of words can play a role in leading the child to understand their meaning. He is simply observing that it is not sufficient for a child to learn to use them. Knowing an association between a word and an object is not the same as knowing what to expect when uttering the word to someone, or as knowing what to do when hearing the word uttered. This suggests that the relation word-object cannot by itself be constitutive of what we call meaning. A variation of the language-game of §2 is introduced in §8 and similar observations about the limitations of ostensive teaching are discussed in §§9-10. We can summarize this idea in the following remark:

Remark 5.1.2 *Explaining a relation between a word and an object is not enough to explain the meaning of the word.*

Although the previous remark concedes that a relation between word and object could be established by ostensive teaching, subsequent observations raise questions regarding the possibility of doing so unequivocally. Consider the example, introduced in §28, of trying to define the number two by pointing to two nuts and saying "That is called 'two'". It is quickly suggested that such an attempt at an ostensive definition might run into trouble from the simple fact that there is always more than one option available for the person trying to interpret the definition:

> But how can the number two be defined like that? The person one gives the definition to doesn't know *what* it is that one wants to call "two"; he will suppose that "two" is the name given to *this* group of nuts! [...] And he might equally well take a person's name, which I explain ostensively, as that of a colour, of a race, or even of a point of the compass. That is to say, an ostensive definition can be variously interpreted in *any* case. (2009, §28, p. 17e)

The following section (§29) explores the possibility that one could clarify the explanation by specifying that one is trying to define a number. This also runs into problems, since if our model of explanation is ostension, then an ostensive explanation of the word 'number' must be given as well. This could then easily lead us into an infinite regress, where to explain what 'number' means, given the underdetermination of ostension, one would require an ostensive explanation of another word, and so forth *ad infinitum*. The point, restated in §30 and illustrated in §§31-32, is that one forgets how much must already be in place for an ostensive definition to work. Establishment of a word-object relation is thus not fully guaranteed by this method. The aim is not to conclude that establishing such a relation is therefore not possible, but simply to remind us that there is always room for misinterpretation when attempting to do so.

The sections that follow (§§33-36) consider a proposal to break this infinite regress by potentially identifying the type of pointing that is being done. The idea is that there would be something characteristic about pointing at a number, or at a shape, or at a color, that could enable one to guess *what* is being defined without needing the recourse to another word in the language. But what would it be like to point at a piece of paper, first pointing at its shape, then its color, then its number? By repeatedly putting this idea into question, Wittgenstein seems to suggest that the belief that this is possible might be somewhat of a myth. Not only is it dubious to think that we can do this by ourselves, but we have to remember that for that to work in an ostensive definition, we have to think of the individual who we want to give the definition to. Whether the definition works will depend on the interpretation she gives, and how she begins to use the word being defined. This, again, is open to mistakes. Wittgenstein wants here to poke more holes in the intuitions that can surround the idea of meaning as correspondence to everyday objects, like the idea that establishing this correspondence is straightforward and error-free. This leads us to our final remark for this section:

Remark 5.1.3 *Establishing a relation word-object suffers from underdetermination.*

The remarks discussed in this section draw attention to the lack of generality and limited explanatory power of externalism as a theory of meaning. If my interpretation is correct, we can thus read §§1-38 of the *Philosophical Investigations* as defending a substantial negative view that one should not conceive of meaning solely as a relation between words and everyday objects in the world.

5.2 Simples

Logical atomism is a theory associated with Russell (1919) and the early work of Wittgenstein (1922). Its perspective regarding meaning, together with the reasoning that can lead to it, is illustrated in the *Philosophical Investigations* in the following passage:

> It can be put like this: *a name ought to really signify a simple.* And one might perhaps give the following reasons for this: the word "Nothung", say, is a proper name in the ordinary sense. The sword Nothung consists of parts combined in a particular way. If they are combined differently, Nothung does not exist. But it is clear that the sentence "Nothung has a sharp blade" has a *sense*, whether Nothung is still whole or has already been shattered. But if "Nothung" is the name of an object, this object no longer exists when Nothung is shattered into pieces; and as no object would then correspond to the name, it would have no meaning. But then the sentence "Nothung has a sharp blade" would contain a word that had no meaning, and hence the sentence would be nonsense. But it does have a sense; so there must still be something corresponding to the words of which it consists. So the word "Nothung" must disappear when the sense is analysed and its place be taken by words which name simples. It will be reasonable to call these words the real names. (2009, §39, pp. 23e-24e)

The characterization of the position is exemplified again in §46 with a quote from Socrates and an explicit reference to Russell and Wittgenstein's early work.[3] The idea is explored from §39 up until §64, where a number of remarks are put forward to question its plausibility.

The versions of logical atomism presented by the two authors may differ slightly[4], but they are both anchored in the following core tenets: the purpose of language is essentially to assert or deny facts; the meaning of a sentence, which is a complex of signs, can be determined by an analysis of its components; when the proposition expressed by a sentence is completely analyzed, it consists only of simple symbols, which have meaning by corresponding to objects; both simple symbols (also called names) and objects are irreducible, in the sense that they cannot possibly be decomposed further. Wittgenstein's attacks on some of these assumptions have already been discussed. In this section, I focus mainly on criticism that is specific to this position.

The sections immediately following the characterization of the position (§§40-45) start by suggesting an alternative to one of the core assumptions of the line of argumentation presented in §39. Wittgenstein goes back to a variation of the language-game of §2, introduced in §15, where building stones can have names. A situation is imagined where a tool named "N" is broken and the builder A sends B the sign "N":

> Now suppose that the tool with the name "N" is broken. Not knowing this, A gives B the sign "N". Has this sign a meaning now, or not? – What is B to do when he is given it? – We haven't settled anything about this. One might ask: what *will* he do? Well, perhaps he will stand there at a loss, or show A the pieces. Here one *might* say: "N" has become meaningless; and this expression would mean that the sign

[3] Wittgenstein quotes a passage from the *Theaetetus* as follows:

> "If I am not mistaken, I have heard some people say this: there is no explanation of the *primary elements* – so to speak – out of which we and everything else are composed; for everything that exists in and of itself can be *signified* only by names; no other determination is possible, either that it *is* or that it *is not*... But what exists in and of itself has to be... named without any other determination. In consequence, it is impossible to give an explanatory account of any primary element, since for it, there is nothing other than mere naming; after all, its name is all it has. But just as what is composed of the primary elements is itself an interwoven structure, so the correspondingly interwoven names become explanatory language; for the essence of the latter is the interweaving of names." (2009, §46b, p. 25e)

It is anachronistic to call the ideas presented in this quote 'logical atomism', since the label was only invented later by Russell to describe his philosophy (see Klement, 2016), but the key ideas of Russell's approach are certainly there in Socrates's reasoning.

[4] See Klement (2016) and Proops (2017).

> "N" no longer had a use in our language-game (unless we gave it a new one). "N" might also become meaningless because, for whatever reason, the tool was given another name, and the sign "N" no longer used in the language-game. – But we could also imagine a convention whereby B has to shake his head in reply if A gives the sign for a tool that is broken. – In this way, the command "N" might be said to be admitted into the language-game even when the tool no longer exists, and the sign "N" to have meaning even when its bearer ceases to exist. (2009, §41, p. 24e)

Here we have another example of a situation where the logical atomist would say that, since the tool supposedly referred to by "N" no longer exists, if we want to say that the sign still has a meaning then it must be in virtue of a correspondence with something else that still exists.

Wittgenstein's approach is to counter this apparent necessity with an alternative. Namely, whether the sign has a meaning or not will depend on whether it has a use or not. Thus we can imagine situations where the name still refers to a tool but one would say the sign no longer has a meaning because it stopped being used in the language-game. Conversely, we can imagine situations where the name refers to a broken tool but one would still say it has a meaning because there are appropriate actions for uses of names when the tools they supposedly refer to are broken. Even names that never named any tool could be said to have a meaning as long as A and B know what to do with them (§42). The remarks in these sections revolve around the importance of considering use when reflecting on meaning. This relates back to the discussion in Section 4.2. With respect to logical atomism, these remarks simply attempt to undermine the necessity of embracing it by presenting another plausible alternative. This is summarized in the following remark:

Remark 5.2.1 *A linguistic expression does not need to refer to existing objects in order for it to be said to have meaning.*

A discussion starting in §47 draws attention to problems with another notion that is essential for logical atomism: the dichotomy between simple and composite. The question that drives it is the following: "What are the simple constituent parts of which reality is composed?" (2009, §47, p. 25e) We are invited to consider several examples of everyday objects, such as a chair, a tree, or a chessboard, and properties of objects, such as color and length, and think of different ways we could consider them as composite and what the simple constituents would be in those situations. A chair could be said to be composed of pieces of wood, or molecules, or atoms (§47a); a tree composed of trunk and branches, or many different sized individual branches (§47c); a chessboard composed of 32 white squares and 32 black squares, or of the colors white and black and a grid (§47d). A color can be considered simple or a combination of other colors; a length could be said to be simple, composed of smaller lengths, or even of a bigger length and another subtracted from it (§47e).

The problem with the driving question is that it does not have a straight answer. Logical atomism assumes the existence of absolute simple constituents of reality. But as soon as we start looking for them, we realize that what one counts as simples depends on the kind of composition one is interested in. Different notions of composite imply different simples, *e.g.* physical decomposition could lead to talk of atoms or quarks, visual decomposition of colors and shapes, and so forth. Wittgenstein's strategy is clear:

> To the *philosophical* question "Is the visual image of this tree composite, and what are its constituent parts?" the correct answer is: "That depends on what you understand by 'composite'." (And that, of course, is not an answer to, but a rejection of, the question.) (2009, §47f, p. 27e)

He is not interested in providing an answer, but in dissolving the driving question, and with it undermine a crucial assumption of logical atomism. The same point is illustrated again in the following sections (§§48-49) by applying the method of §2, *i.e.* devising a toy language-game where the idea seems to correctly apply and drawing attention to the problematic hidden assumptions. The conclusion is the same, which we can summarize in the following remark:

Remark 5.2.2 *Different notions of composite yield different kinds of simples, thus it makes no sense to speak of simples in the absolute.*

Further intuitions regarding the nature of simples are dissected in §§50-59. This is accompanied by a proposal of an alternative way of thinking about them. The first idea challenged is that one can attribute neither being nor non-being to an element[5]. If an element is something that, by definition, is not composite, it cannot have different properties that characterize it, otherwise it could be said to be composed of those properties. Therefore, it seems, an element can only be named, not described. Thus it would make no sense either to say of an element that it exists or that it does not exist, since the mere use of its sign would guarantee, by virtue of reference, that assertions of its existence are tautological and statements of its non-existence nonsensical.

[5] 'Simple' and 'element' are used interchangeably.

Wittgenstein seems to consider this line of reasoning circular, for example in the following passages:

> What does it mean to say that we can attribute neither being not non-being to the elements? – One might say: if everything that we call "being" and "non-being" consists in the obtaining and non-obtaining of connections between elements, it makes no sense to speak of the being (non-being) of an element; just as it makes no sense to speak of the destruction of an element, if everything we call "destruction" lies in the separation of elements. (2009, §50, pp. 28e-29e)
>
> "A *name* signifies only what is an *element* of reality – what cannot be destroyed, what remains the same in all changes." – But what is that? – Even as we uttered the sentence, that's what we already had in mind! (2009, §59, p. 33e)

The properties attributed to the elements are baked into their postulation, and the plausibility of their existence is dependent on a conception of meaning as correspondence. If we take that away, we would perhaps not be at all compelled to accept logical atomism as a credible or useful theory of meaning.

Wittgenstein proposes an explanation for why one might be tempted to think of simples in the terms we just described. He considers likely candidates for simples, including the standard meter in Paris and a color sample for sepia (§50b). Of the former, it does not make sense to say that it is or is not 1 meter long, for it is by virtue of comparisons to it that we speak of lengths and distances in meters. The same happens about whether a sepia sample is or is not itself sepia. These are, however, not metaphysical pronouncements about special entities. They are simply grammatical remarks about the role these samples play in those particular language-games. The standard meter works as a reference point, or as Wittgenstein calls it a paradigm, in language-games that involve communicating measures in meters. Its status is special in that "it is not something that is represented, but is a means of representation." (2009, §50c, p. 29e) The reason why a paradigm seems logically indestructible is that we cannot do away with it without doing away with the whole language-game. This is, again, only a feature of the particular language-games where these paradigms serve as instruments to enable us to make moves in those games. They are thus contingent to them and not general properties of metaphysical objects, as logical atomism would lead us to believe. The considerations relevant to this section can be summarized in the following remark:

Remark 5.2.3 *What logical atomism sees as symbols referring to metaphysical simples can alternatively be seen as linguistic expressions that have a contingently special role in particular language-games.*

The observations in §§60-64 address the methodological partner of logical atomism: the method of analysis. Logical atomism sees analysis as the way to reveal the meaning of linguistic expressions. As such, it sees an analyzed name as better than its un-analyzed form. To start questioning this assumption, an example of a broom is put forward. If we consider the broom to be composed of a broomstick and a brush, should we say that a statement containing the word 'broom' is *actually* about the two simpler components? The example is not accurately representative of the ideas of logical atomism, since a broomstick and a brush are not good examples of simples, but it serves to illustrate the underlying idea that language stands in need of analysis, *i.e.* that in order to understand a sentence like "My broom is in the corner" one needs to break down the complex meaning of 'broom' into smaller components.

Against this, Wittgenstein simply counters with two objections. The first is an appeal to everyday intuitions:

> Then does someone who says that the broom is in the corner really mean: the broomstick is there, and so is the brush, and the broomstick is fixed in the brush? – If we were to ask anyone if he meant this, he would probably say that he had not specially thought of either the broomstick or the brush. And that would be the *right* answer, for he did not mean to speak either of the stick or of the brush in particular. Suppose that, instead of telling someone "Bring me the broom!", you said "Bring me the broomstick and the brush which is fitted on to it!" – Isn't the answer: "Do you want the broom? Why do you put it so oddly?" — Is he going to understand the further analysed sentence better? – This sentence, one might say, comes to the same thing as the ordinary one, but in a more roundabout way. (2009, §60, pp. 33e-34e)

What we're being asked to consider is whether, when using a word that supposedly refers to something complex, we have any intuition of having intended to refer to its components, or conversely whether, when hearing a sentence where the components are explicitly broken down, we would understand it better in that form. The suggestions are that one does not typically think of the components of a supposed complex when producing a sentence about that complex, and that a sentence where these components are explicitly mentioned is not more understandable, and might actually appear more odd and unnecessarily complicated. It runs counter to these intuitions to think that there is any need for analyzing a complex to get to a hypothetical *actual* meaning. This is not to claim that an analysis is not possible, but simply that it is often neither necessary nor useful.

The second objection builds on this to question whether a supposed analysis produces an equivalent, but somehow more fundamental, representation of the same thing. We are asked to imagine two language-games, one where complex objects have names and another where only the constituents of the objects are given names (§60). The question immediately arises of whether these two games can be said to be two different forms of the same game, and how we would go about defining an equivalence between statements in both. Although one could perhaps say that a given statement in one language-game "comes to the same thing" (2009, §61, p. 34e) as a particular statement in the other, this does not mean that the two statements are fully interchangeable. We might not be able to use the unanalyzed form in the game where only constituents have names, neither the other way around. If that is the case, each game has its merits, and none is more fundamental than the other:

> We may think: someone who has only the unanalysed form has got it all. – But can't I say that an aspect of the matter is lost to the latter no less than to the former? (2009, §63, p. 35e)

This is illustrated again in §64 with a variation of the game of §48, where rather than monochrome squares one could have squares with two different colors on each half[6]. The conclusion is the same: names for such multi-colored squares would not stand in need of analysis, nor would it be necessarily possible to replace one game by the other. The suggestion is that the possibility of analyzing a name does not necessarily imply that an analyzed form is hidden when we use the unanalyzed form, we might just be playing a different language-game.

These considerations about analysis are relevant to the intuitions behind logical atomism since they chip away at the idea that linguistic expression can or need to be broken down in a unique way that reveals their fundamental constituents. Wittgenstein's remarks remind us that, first, one can understand a word like 'broom' without analyzing it (often even better), and second, that there is no unique way of performing such analysis independently of the particular context. These points make it less appealing to think that there is such a thing as a unique meaning of a linguistic expression to be found in the simple elements it corresponds to. Further evidence that this is what Wittgenstein has in mind can be found, for example, in the

[6] See Section 3.2 for a more in-depth discussion of this example.

passage from the *Blue Book* quoted in Section 3.1 (Wittgenstein, 2002, pp. 27-28). We can summarize these points in the following remark:

Remark 5.2.4 *Analyzed forms of linguistic expressions are neither better nor unique.*

These remarks on logical atomism suggest that Wittgenstein came to reject his own older theory as not especially plausible or useful. Furthermore, they seem to dismiss the urge to sublimate the discussion of meaning to the realm of the metaphysical by anchoring the discussion in more everyday contexts. One can thus read §§39-64 as a rejection of the idea that linguistic expressions stand in need of analysis for their meaning to be unearthed, and that they get their meaning by being composed of irreducible elements that correspond to simple objects in the world.

5.3 Mental entities

Another common intuition about language is the idea that words get their meaning by standing in relation to entities in the mind, be it concepts, ideas, feelings, sensations, or other inner private experiences. Wittgenstein explores some notions related to this view mostly in §§243-315, in large part by reflecting on a related problem: the possibility of a private language. This, he characterizes as follows:

> But is it also conceivable that there be a language in which a person could write down or give voice to his inner experiences – his feelings, moods, and so on – for his own use? — Well, can't we do so in our ordinary language? – But that is not what I mean. The words of this language are to refer to what only the speaker can know – to his immediate private sensations. So another person cannot understand the language. (2009, §243b, p. 95e)

Following Stern (2004, p. 174), I will call the type of privacy here described 'super-privacy'. A super-private language is one in which expressions refer *exclusively* to entities that are available only to the speaker, and nothing else. If meaning is constituted by these entities, this implies that nobody else could ever know the meaning of the expressions that make up such a language. Discussion of this idea is relevant because it is warranted by a conception of meaning as correspondence to mental entities. I will henceforth call this *internalism*[7]. Debating the possibility of a

[7] This term is used in contemporary philosophy in the context of different topics, with correspondingly various definitions. As with my use of the term externalism, it would be anachronistic to portray Wittgenstein as attacking those positions. I use the term simply to label the general intuition that underlies a particular picture of meaning.

super-private language allows one to consider internalism in a scenario that excludes additional issues related to the use of language as a means of communication. The content of these sections of the *Philosophical Investigations* has often been called the 'private language argument'.

Wittgenstein starts the discussion by reflecting on how something like a super-private language could be taught and learned. The word 'pain' is used as a paradigmatic example of a word for which it is easy to think that its meaning lies exclusively in its correspondence to a super-private experience. However, such a word needs to be taught and learned. How would it be possible to teach a child the meaning of the word 'pain' if the meaning were constituted only by his super-private experience? How would an adult refer to an experience that is not available to them in order to establish the correspondence with the word? What typically happens, it is suggested in §244, is that teaching is done by observing the child's behavior, making a guess (one presumes) that the child is in pain, and teaching her new "pain-behavior", *i.e.* using a word to signal this rather than crying. Wittgenstein is not here adumbrating a fully behavioristic theory of language learning (he himself seems to reject this in §§304-308); he is merely reminding us of some non-private dimensions of what can, at first glance, be considered as super-private as it gets.

Behavior not only enables us to teach and learn these words and expressions that purportedly refer to super-private experiences, but is also very important in sustaining our beliefs regarding those very experiences. This is defended in §§281-288, with the main point summarized in the first passage:

> "But doesn't what you say amount to this: that there is no pain, for example, without *pain-behaviour*?" – It amounts to this: that only of a living human being and what resembles (behaves like) a living human being can one say: it has sensations; it sees; is blind; hears; is deaf; is conscious or unconscious. (2009, §281, p. 103e)

The following sections explore different examples to draw attention to the importance of identifying or projecting some kind of human behavior in order to accept attributions of the ability to have private sensations. A pot in a fairy tale can be imagined to see and hear, but we also attribute it the ability to speak (§282). Although it is difficult to ascribe private sensations to an inanimate object or a corpse, we do not have the same qualms with attributing pain to a fly as soon as it starts wriggling (§284). Speaking, wriggling, making facial expressions (§285), these are all behaviors that naturally pair up with ascriptions or explanations of pain and other private sensations. Behavior, it is later said, belongs to the language-game (§300). The analogies in these sections serve to remind us of the role that behavior plays in our understanding of what pain is, and thus of what the word 'pain' means. This contributes to undermining the picture of meaning as depending only on words and mental entities. This leads us to the first remark on this topic:

Remark 5.3.1 *There are non-private dimensions relevant to the meaning of expressions that supposedly correspond to super-private experiences.*

Without the public dimension of behavior, one would require other criteria to underwrite our practices of learning, explaining, and understanding meaning. This is mostly explored in §§253-292. Wittgenstein starts with the problem of how to internally identify super-private experiences:

> "Another person can't have my pains." – *My* pains – what pains are they? What counts as a criterion of identity here? Consider what makes it possible in the case of physical objects to speak of "two exactly the same": for example, to say, "This chair is not the one you saw here yesterday, but is exactly the same as it". [. . .]
> I have seen a person in a discussion on this subject strike himself on the breast and say "But surely another person can't have THIS pain!" – The answer to this is that one does not define a criterion of identity by emphatically enunciating the word "this". Rather, the emphasis merely creates the illusion of a case in which we are conversant with such a criterion of identity, but have to be reminded of it. (2009, §253ac, p. 97e)

This is relevant for the idea of a super-private language if one imagines that such a language could get started by associating words and super-private experiences in a form analogous to an ostensive definition (§258): one concentrates his attention on such an experience and commits to memory the connection with a word. However we do this, the purpose would be to later access the particular experience when hearing the word, or to produce the appropriate word when wishing to refer to the experience.

But how do we know we have performed these steps correctly? Aren't we prey to confabulation, to simply convince ourselves that we have done the right thing even if we haven't? This problem is illustrated with examples of mental acts like consulting a mental image of a timetable to determine the schedule of a train (§265), looking at a clock in one's imagination to tell the time (§266), or making imaginary load tests on a projected bridge to justify the choice of dimensions in a design (§267). In all of these cases one might be under the illusion that something has been achieved, but a great deal of doubt would remain before external validation is performed. Beyond these sections we also find the examples of the hypothetical scenarios of talking to oneself without ever having spoken an audible language (§344) and learning to calculate in the head without ever calculating aloud or on paper (§385). These scenarios sound at least odd if not impossible to imagine, exactly because the possibility of validating those mental acts is inaccessible even for the person hypothetically conducting them.

This issue casts doubt on the coherence of the whole idea of creating a super-private language by a process analogous to ostension. In §269, Wittgenstein makes a rough distinction between three types of ways to relate to linguistic expressions: not understanding, thinking one understands, and understanding correctly. A super-private language of an individual would consist of sounds that no one else understands but that he appears to understand. One could never say that he understands the language correctly, for the lack of external criteria that could validate his beliefs regarding how to use those sounds. A super-private language, if possible, is therefore not on par with the natural languages we know. This is important, because exactly what makes the latter different is their public dimension which provides the criteria that enable us to talk about knowing the meaning of a linguistic expression rather than merely believing one knows it. This might sound like a metaphysical or essentialist pronouncement about what a language is or isn't, but that is not how it is intended. We can recall an earlier passage in order to clarify this:

> The criteria which we accept for 'fitting', 'being able to', 'understanding', are much more complicated than might appear at first sight. That is, the game with these words, their use in the linguistic intercourse that is carried on by their means, is more involved – the role of these words in our language is other than we are tempted to think. (2009, §182b, p. 79e)

The same idea is restated in a later passage in §573. The defense of the need for external criteria is a reminder of how we talk about knowing a language, knowing the meaning, even of expressions that supposedly refer to private experiences, and how having external criteria is important in those language-games.[8] These considerations again support the view that a postulated mental entity that corresponds to a word like 'pain' cannot constitute the meaning of the word by itself. We can summarize this in the following remark:

Remark 5.3.2 *Linguistic expressions referring to private experiences require the existence of public criteria.*

[8] With this in mind, we can understand the often quoted passage declaring that "[a]n 'inner process' stands in need of outward criteria" (2009, §580, p. 161e) not as a statement about inner processes as entities, but as a reminder about how one usually uses the expression 'inner process'.

In the passages discussed so far, Wittgenstein seems to merely attempt to complement internalism by suggesting that behavior in particular, and public criteria in general, are important additional dimensions of what we typically call the meaning of an expression of a private experience. But there are also other passages that seem to more fully reject private experiences as even contributing to the meaning of private expressions. Consider the following example: "Imagine a person who could not remember *what* the word 'pain' meant – so that he constantly called different things by that name – but nevertheless used it in accordance with the usual symptoms and presuppositions of pain." (2009, §271, p. 101e) Rather than taking this as a challenging scenario, Wittgenstein quickly dismisses it by stating that this is how we all use the word 'pain'. The private experience of remembering or not the meaning seems to be considered irrelevant *as long as* there is an external criterion available or an accordance of behavior. The following sections (§§272-280) illustrate this point by delving into the use of color terms like 'red' and 'blue'.

Slightly later on we encounter a famous passage in the *Philosophical Investigations* where Wittgenstein lays out the often-called beetle-in-the-box thought experiment:

> Well, everyone tells me that he knows what pain is only from his own case! — Suppose that everyone had a box with something in it which we call a "beetle". No one can ever look into anyone else's box, and everyone says he knows what a beetle is only by looking at *his* beetle. – Here it would be quite possible for everyone to have something different in his box. One might even imagine such a thing constantly changing. – But what if these people's word "beetle" had a use nonetheless? – If so, it would not be as the name of a thing. The thing in the box doesn't belong to the language-game at all; not even as a *Something*: for the box might even be empty. – No, one can 'divide through' by the thing in the box; it cancels out, whatever it is.
>
> That is to say, if we construe the grammar of the expression of sensation on the model of 'object and name', the object drops out of consideration as irrelevant. (2009, §293bc, pp. 106e-107e)

The scenario described has been variously interpreted[9]. It can be construed as another application of the "method of §2", in that the narrator designs a situation where the interlocutor's views seem to apply correctly, only to let us see that, under close inspection, it does not work as expected.

[9] See Stern (2007) for an overview.

In this case, the issue is the same as was adumbrated in §§272-280: a publicly shared word or expression cannot get its meaning from a super-private experience since, given that we cannot access other people's experiences, these could vary wildly and we could never be sure about what others meant. It is important to note the conditional formulation of §293c. Between §294 and §315, Wittgenstein makes it clear that he does not deny that private experiences exist or inner processes can accompany or even trigger our use of words. He is simply stating that, *if* we conceive of these experiences as super-private they cannot constitute the meaning of the expressions that supposedly refer to them, because they would be irrelevant as an explanatory factor. The following remark summarizes this point:

Remark 5.3.3 *Public words and expressions cannot get their meaning from super-private entities or experiences.*

There is an additional aspect of Wittgenstein's attack on internalism: the suggestion that its intuitions are rooted in grammatical confusion. Sections §§246-252 begin to attempt a deconstruction of some assumptions about super-private entities. The first sections articulate the idea that knowledge of these entities or experiences is qualitatively different between the person that experiences them and others. Namely, whereas when one is in pain one certainly knows it, when one thinks another person is in pain there is always room for doubt, *i.e.* there is always a possibility that the other person is mimicking the external behavior without having the internal experience. Hence, "it makes sense to say about other people that they doubt whether I am in pain; but not to say it about myself." (2009, §246c, p. 96e) Without denying this intuition, Wittgenstein's subtle remarks up to §252 attempt to warn against potential misinterpretation of this kind of statements.

For example, the observation just quoted can lead us to state that pain is thus private. There is nothing wrong with such a statement by itself. The problem only arises if we take it as a statement about an entity, and imagine that the word 'pain' refers to it. This touches on two earlier remarks: that we should be careful not to "confound the meaning of a name with the *bearer* of the name" (2009, §40, p. 24e) or, one could add, even assume that because there is a name there is a bearer; and that some sentences, despite looking like they are about something external to them, are actually merely about the use of the words or expressions involved (see also §58). The cryptic §248 is an example that is meant to evoke this idea: just as "One plays patience by oneself" shouldn't be seen as an empirical statement about the game of patience as an external entity, but rather as part of setting up the game, so should "Sensations are private" be interpreted as setting up a language-game that uses the word 'sensation', rather than as a statement about an independent thing referred to by that word.

Later one gets a somewhat more explicitly portrayal of this idea:

> What does it mean when we say, "I can't imagine the opposite of this" or "What would it be like if it were otherwise?" – For example, when someone has said that my mental images are private; or that only I myself can know whether I am feeling pain; and so forth.
> Of course, here "I can't imagine the opposite" doesn't mean: my powers of imagination are unequal to the task. We use these words to fend off something whose form produces the illusion of being an empirical proposition, but which is really a grammatical one. (2009, §251ab, p. 96e)

What is suggested is that, for example, the sentence "Only I can know whether I am feeling pain", although looking like an empirical proposition about pain, is actually more a statement about the words in the sentence than a statement of empirical fact, just like "One plays patience by oneself" (2009, §248, p. 96e), "Every rod has a length" (2009, §251d, p. 97e), or "This body has extension" (2009, §252, p. 97e). These are sentences that set up the rules of particular language-games and this is why we either reply "Of course!" or "Nonsense!", we either reject or accept playing the game.

The issue returns again later, interspersed with the beetle-in-the-box thought experiment. In §§294-308, Wittgenstein addresses potential concerns that he is denying mental processes and the charge of being a behaviorist. To the accusation that he repeatedly reaches the conclusion that the sensations are a Nothing, he replies:

> Not at all. It's not a Something, but not a Nothing either! The conclusion was only that a Nothing would render the same service as a Something about which nothing could be said. We've only rejected the grammar which tends to force itself on us here. (2009, §304, pp. 108e-109e)

The grammatical confusion is to believe that there is necessarily a Something because we talk as if we are referring to something. This is what Wittgenstein means when saying that the grammar is forcing itself on us. When we say "my table" we think of it as referring to an external object, so when we say "my pain" we can think we must be referring to some entity as well, hence creating the image of a private experience as a thing, albeit internal rather than external. The grammatical similarity invites us to make the analogy without even realizing it.

Replying to a further accusation (by the interlocutor's voice) that he considers everything except human behavior a fiction, he retorts that he is only talking about a *grammatical* fiction (§307), and spells out how these fictions come about in the next section:

> How does the philosophical problem about mental processes and states and about behaviourism arise? — The first step is the one that altogether escapes notice. We talk of processes and states, and leave their nature undecided. Sometime perhaps we'll know more about them – we think. But that's just what commits us to a particular way of looking at the matter. For we have a certain conception of what it means to learn to know a process better. (The decisive movement in the conjuring trick has been made, and it was the very one that seemed to us quite innocent.) (2009, §308, p. 109e)

The grammatical confusion is again to take the existence of expression like 'mental process' or 'mental state' as implying the existence of corresponding entities. Wittgenstein, as he repeatedly states, does not want to deny the existence of mental life or experiences, only to expose this conjuring trick and suggest that, just because we have words like 'pain' or 'sensation' that seem to relate to something super-private, that does not necessarily imply that there are mental correlates of these words which constitute their meaning. We might just be being misled by an unreflected analogy between empirical and grammatical propositions. We can summarize this point in the following remark:

Remark 5.3.4 *The existence of mental entities corresponding to private expressions is not guaranteed by the existence of these expressions in our language.*

These remarks taken together raise doubts about internalism as a useful or even coherent theory of meaning. The only way out seems to be to "make a radical break with the idea that language always functions in one way, always serves the same purpose: to convey thoughts – which may be about houses, pains, good and evil, or whatever." (2009, §304, p. 109e) It would be good to abandon the picture of meaning as a form of correspondence, and the idea that linguistic expressions have meaning by corresponding to other entities. In the following section, I reflect on whether or not the framework of signaling games gives us a way of doing so.

5.4 Correspondence in signaling games

I presented the remarks in this chapter as addressed against three particular variants of the correspondence picture of meaning. Pointing out issues with these variants is not, I believe, an invitation for the development of another variant that would not suffer from them. It is rather a way of rejecting the general picture. Some remarks are specific, but many apply quite generally. What is the correspondence picture of meaning? We can describe it again, by slightly paraphrasing a quote from §1b, as the following idea: every linguistic expression has a meaning; this meaning is correlated with the linguistic expression. I use 'linguistic expression' to encompass what we usually call 'signs', 'words', or 'sentences'. Externalism, logical atomism, and internalism are particular variants of this picture that differ in how they see the nature of this correlate. Even if we prefer one variant over another, the basic idea they share seems intuitive and innocuous. What is Wittgenstein's problem with this picture?

Although it might be somewhat obvious, I think it is important to point out that it is very unlikely Wittgenstein's criticism of this intuition is aimed at correcting our everyday use of it. In this respect, it seems safe to take him at his word when saying that "[p]hilosophy must not interfere in any way with the actual use of language" (2009, §124, p. 55e). Wittgenstein is not trying to reform language (§132), at least not in how it is used in an everyday non-philosophical context. The criticism of the correspondence picture of meaning is aimed at attempts at theorizing about language that take the idea for granted.

One reason to find the correspondence picture problematic in philosophical theorizing has to do with the remarks discussed in Section 3.1. The idea that a linguistic expression can have a meaning that corresponds to it motivates seeing meaning as a separate entity which, given that we do not have any direct access to it, is hidden. This can lead to the expectation that linguistic expressions have some single final completely analyzed form that can be unearthed through an investigation. This kind of project is, according to Wittgenstein, misguided (Remark 3.1.2, p. 59). Talk of meaning only in terms of correspondence to linguistic expressions additionally encourages ignoring all the aspects that make meaning vary, like context, purpose, and the role of agents in how they make use of those expressions. It can invite hypostatizing meanings as entities, something Wittgenstein would also be against (Remark 3.1.3, p. 61).

Other reasons to take issue with the correspondence picture of meaning can be generalized from some of the specific remarks addressed at its variants. One of the most central is a recurrent theme in Wittgenstein's attempts at the dissolution of various apparent problems: the issue of heterogeneity. As mentioned in Section 3.2, one can see the search for a theory of meaning as an attempt at defining the word 'meaning'. One problem immediately faced by such an enterprise is that we use the word in various ways. Even if we ignore uses that do not clearly relate to linguistic meaning[10], there are kinds of signs, words, and sentences that we say are meaningful but might not fit the correspondence picture very well. Although Remark 5.1.1 (p. 112) was discussed in the context of externalism, it is formulated in a way that applies to the general idea. That linguistic expressions have meanings which are correlated with them might seem intuitive for common nouns, names, sentences that we can paraphrase in a simpler way. And that is in the context of particular language-games. Trying to generalize this picture to the whole of what we call language, given its heterogeneity and dynamism (Remark 4.1.2, p. 93), might not be warranted. The uses of the word 'meaning' are tied together by family resemblance, and meaning is just not something we can define in all generality.

Further criticism of the three variants of the correspondence picture raises concerns that, even if we restrict it to scenarios where it seems to apply somewhat innocuously, it might be a limited notion. The imaginary builder does not use the word "slab" solely to evoke images in the mind of the assistant (Remark 5.1.2, p. 113), understanding an analyzed linguistic expression is not necessarily sufficient to understand its unanalyzed counterpart (Remark 5.2.4, p. 122), and there is more to the meaning of a word like "pain" than some correlated super-private experience (Remarks 5.3.1, p. 124, and 5.3.3, p. 127). This applies to the correspondence picture in general:

> If we say, "Every word in the language signifies something", we have so far said nothing *whatever*; unless we explain exactly *what* distinction we wish to make. (2009, §13, p. 10e)

Saying that every linguistic expression has a meaning is, by itself, a vacuous claim. It can serve a purpose if we are trying to distinguish a word like "table" from a sequence of letters like "tnetennba". Explaining the meaning of a linguistic expression in terms of an object, a paraphrase, a concept, or anything else, can help prevent or resolve a misunderstanding, but only if everything else is already in place (Remark 3.2.1, p. 67). Defining the king in chess by establishing a correspondence with the piece does not in general determine how the piece is to be used in the game (§31). Similarly, establishing the meaning of an expression by correspondence is not sufficient to say everything about its meaning.

[10] For example, uses like "Life has meaning" or "She means a lot to me".

A demand that usually accompanies the correspondence picture is that meaning must be determinate. Against this, Wittgenstein points out the problems of the underdetermination of ostensive definitions. If one can never establish an exact correspondence between a word and an everyday object by physical ostension (Remark 5.1.3, p. 115), or with a super-private experience by some mental equivalent of that (Remark 5.3.2, p. 125), how is one supposed to ever fix those correlates as their meaning? A similar issue could be raised regarding sentences and their analyses or paraphrases, given how those are tied to language-games, and the latter are numerous and subject to change. The problem of underdetermination and its implications for the correspondence picture of meaning appear more prominently in Wittgenstein's remarks about rules and rule-following (see Chapter 6). For now, it suffices to say that the points made about the determinacy of meaning in the context of externalism and internalism generalize well to the correspondence picture of meaning.

In Section 4.3, I argued that the signaling games framework allows us to study meaning by focusing on use. They therefore permit us to forego the intuitive urge to see meaning in terms of correspondence, if we take Wittgenstein's remarks on use as a methodological advice rather than as a theory of meaning, as I defended in the discussion leading up to Remark 4.2.2 (p. 100). This is, to a certain extent by design, as evidenced by the following passage:

> Communication by conventional signals is a commonplace phenomenon, so much so that we must make an effort not to take it for granted. We could exercise our tacit understanding all we want without ever making it more explicit. That is what would happen if we started by saying that actions are signals when we endow them with meanings. This truism will bring us no nearer to describing the phenomenon of signaling without depending on our prior tactic understanding thereof. So let us describe the phenomenon in other terms and leave meaning to look after itself. (Lewis, 1969, p. 122)

I believe this attitude is in line with the Wittgensteinian remarks discussed so far. Lewis' suggestion seems to be to forget our tacit understanding of signals as being endowed with meaning, describe them in other terms, and we'll understand meaning better. The fits the idea of rejecting the correspondence picture of meaning and following Wittgenstein's methodological advice to "look at the use".

Lewis follows this up with the characterization of the Paul Revere signaling game

(Lewis, 1969, pp. 122-125). This is accompanied by the following observation[11]:

> I have now described the character of a case of signaling without mentioning the meaning of the signals: that two lanterns meant that the redcoats were coming by sea, or whatever. But nothing important seems to have been left unsaid, so what has been said must somehow imply that the signals have their meanings. (1969, pp. 124-125)

Again, the attitude just discussed seems to be reiterated. After describing the use of signals, "nothing important seems to have been left unsaid". A thoroughly Wittgensteinian attitude would be to conclude "we are, therefore, done". But the intuitive appeal of the correspondence picture is strong. And although the signaling games framework allows one to avoid it, it does not fully eliminate it. Some authors have therefore tried to provide definitions of the meaning of signals in signaling games that to me go awry by regressing into viewing meaning in terms that are close to the correspondence picture.

Lewis is, surprisingly (given the aforementioned passages), the first one to go down that road:

> I have been trying to demonstrate that an adequate account of signaling need not mention the meanings of signals—at least, not by name. But of course signals *do* have meanings. (Lewis, 1969, p. 143)

He subsequently develops an account of meaning anchored in his notion of convention. Consider the signaling system in Figure 2.1. Let us call σ and ρ the sender strategy and the receiver strategy, respectively, depicted therein. Lewis claims (1969, pp. 143-152) that there are a number of things that we can say about the meaning of, for example, signal m_a in that situation[12]:

1. m_a is a conventional signal in (σ, ρ) that t_1 holds or m_a conventionally means in (σ, ρ) that t_1 holds;

2. m_a is a conventional signal in (σ, ρ) to do a_1 or m_a conventionally means in (σ, ρ) to do a_1.

These are two alternative ways to "give the meaning" of a signal in a signaling system. Lewis makes further distinctions and definitions regarding meaning, which I believe are best addressed after reviewing Wittgenstein's remarks on rule and rule-following (see Chapter 6).

[11] Also quoted in Section 4.3.

[12] These are close paraphrases, merely adapted to use more the notation used in the examples in this thesis.

This conceptualization of signal meaning is likely to align well with our everyday intuitions. Given all the regularity and common knowledge in a signaling system anchored in convention, it can seem natural to want to say something along the lines of "m_a means t_1" or "m_a means a_1". And Lewis' definitions give us a way of saying such things in the context of the signaling games framework. However, they are technical notions. As such, in order to be faithful to the framework, meaning needs to be hedged as "conventional meaning" and the particular sender and receiver strategies that constitute the equilibrium need to be included in the definitions. With these caveats in place, I don't think they are problematic as technical notions, partly because the don't bring anything new; they are mere paraphrased descriptions of the sender and receiver strategies.

Nevertheless, by attempting to interface between the technical notions of the framework and our natural ways of speaking about meaning, one can invite a perspective that is very different from the original objective of leaving "meaning to look after itself." (1969, p. 122) Huttegger (2007b), while revisiting some of the ideas introduced by Lewis, states the following:

> As long as no signaling system or convention is established in a population, signals have no meaning. [...] On the other hand, meaning may be considered as a property of signals in equilibrium. If almost all individuals play according to a signaling system, then signals are representations of parts of the world and have these parts as contents. To be more specific, signals in a signaling system refer to a state of the world and to an act that is a proper response to this state. We will say that signals in signaling systems refer to state-act pairs. (2007, p. 413)

What is described in this passage is a notion of meaning in signaling games that is an instance of the correspondence picture. In equilibrium, signals are said to have meanings (as a property or content) which are constituted by the state-act pairs they refer to. Rather than illuminating the notion of meaning by describing the phenomenon in different terms, we have come full circle and are back to the standard quasi-metaphysical truisms couched in our tacit understanding that Lewis claimed to want to avoid.

Further problems with shifting the approach to meaning in signaling games from the former attitude to the latter, can be realized by taking Wittgenstein's remarks into account. The first thing to note is that it is a very limited account, even if we take it as a notion of meaning within the signaling games framework. The definitions given above attribute meaning to signals only when they are used in equilibrium and the strategies are pure-strategy equivalents[13]. Perhaps it is easy to find what the candidate for a signal's meaning is when both sender and receiver are in a stable situation by consistently (100% of the time) either using the same signal for one and the same state or performing the one and the same action for a given signal. But this makes it unclear at which point in an adaptive process signals "get" their meaning (see Section 4.3), signals can enable coordination even when strategies are not in stable equilibrium (*e.g.* Wagner, 2012) and there are games with stable equilibria that do not consist of pure-strategy equivalents, such as partial pooling in simple Lewis signaling games (*e.g.* Huttegger, Skyrms, Smead, et al., 2009), vague signal use in sim-max games (*e.g.* Franke, Jäger, and van Rooij, 2011; O'Connor, 2014b; Franke and Correia, 2018), and hybrid equilibria in costly signaling games (*e.g.* Huttegger and Zollman, 2016). Strict notions of meaning, like those proposed by Lewis and Huttegger, exclude all of these cases as meaningful signal use.

The matter only gets worse if one considers these notions of meaning as something that can be generalized to the broader context of natural language. Language is deeply dynamic, so it is (almost) never in a fully stable state, and heterogeneous, so each word is likely used or potentially usable in more than one language game (Remark 4.1.2, p. 93). If this picture is correct, notions of meaning that only work for pure-strategy equivalents at equilibrium and nothing else are irrelevant for natural language. This is important to mention, since Lewis himself makes close parallels (1969, pp. 152-159) between his definitions of meaning and Grice's notion of non-natural meaning (Grice, 1957). The latter, however, is supposed to have a much broader applicability. That Lewis finds a comparison between the two a legitimate step to take is perhaps a good demonstration of how choosing to talk about meaning in correspondence terms, rather than talking about use alone, invites hasty generalizations.

I suggested that these definitions of meaning in signaling games were introduced driven by the appeal of the intuitiveness of the correspondence picture. But they might have another source. One can feel a certain urge to want to distinguish a random strategy from a signaling system in terms that go beyond total expected utility and capture something related to communication. Skyrms (2010, pp. 33-47) proposes that we do this using notions from information theory[14]. In particular, in a signaling game model, one can calculate the amount to which the use of a signal by the sender makes it more likely that a certain state obtains. Conversely, the

[13] A pure sender univocally assigns a choice (signals for senders, actions for receivers) per choice point (states for senders, signals for receivers). The equivalents in mixed and behavioral strategies would be, respectively, a strategy that is 100% composed of a single pure strategy, and a strategy that, per choice point, assigns a unique choice with 100% probability.

[14] See Cover and Thomas (2006) for an introduction.

receiver strategy makes it the case that each signal affects, by a certain amount, the probability that a certain action is performed. Omitting some technical details, these are called the signal's amount of information about a state and amount of information about an action, respectively. These values can be used to define a signal's *informational content about states* (resp. *about actions*) as a vector indicating the amount of information it is said to carry about each of the states (resp. actions). Averaging over this vector gives the *quantity of information* in a signal. For example, in a binary signaling game with equiprobable states, like the example discussed in Section 2.1, if each signal is used 50% of the time for each state, that signal carries no information about any state; the amount of information is maximal when the sender strategy perfectly discriminates the state by using a separate signal for each separate state 100% of the time.

One attractive aspect of this notion of quantity of information is that it is a matter of degree. The measures can be calculated for any pair of sender and receiver strategies in a signaling game model. It is neither required that they are pure-strategy equivalents nor that they are in equilibrium. Its applicability is therefore much broader than the notions of meaning suggested by Lewis (1969) and Huttegger (2007b). One can, for example, use it to observe how the amount of information increases gradually as strategies are driven towards a signaling system by an adaptive dynamic in a binary signaling game (Skyrms, 2010, p. 40). Another advantage is that it is a technical notion that captures part of what we feel compelled to say about the differences between how signals are used in random strategies versus separating strategies that foregoes the title of 'meaning'. By introducing it in such a way, Skyrms avoids the invitation for unwarranted generalizations. Quantity of information is not the same as meaning, even though it seems to capture aspects of it. But the separation helps avoid confusion between a technical notion defined within the context of signaling games and an everyday expression used in a broad family of cases.

Despite these advantages, the notion of information can still be seen as problematic since it is presented in terms similar to the correspondence picture of meaning. Skyrms describes signals as *having* informational contents, or as *carrying* information. These ways of speaking seem to elicit misleading analogies, something Wittgenstein also warned us to avoid[15]. A signal is not a container that can be pried open for access to its contents, nor can a signal hand over information like a person handing over a bag of groceries she is carrying. These analogies could encourage one to think, much like in the correspondence picture, that a signal has information, one

[15] See the discussion leading up to Remark 3.1.2 (p. 59).

need only to know how to access it. This apparently leaves all the other elements of the signaling game behind. But looking more closely into the technical definition of information, reveals that this is an issue of presentation.

Following the notation used in Section 2.1, the amount of information a message m carries about a state t is given by the following formula:

$$I(m,t) = \log\left(\frac{P(t|m)}{\Pr(t)}\right)$$

Here, $P(t|m)$ is the probability that state t holds when message m is used, and $\Pr(t)$ is the prior probability that state t holds. According to Skyrms (2010, p. 35), $P(t|m)$ can be given, by Bayes' theorem, as follows:

$$P(t|m) = \frac{P(m|t)}{P(m)}$$

Additionally, the probability that message m is used given the state, *i.e.* $P(m|t)$, is defined by the sender strategy σ. The probability that a message is used simpliciter, *i.e.* $P(m)$, is also defined by the sender strategy, but additionally taking into account the probability that message m is sent in all other states as well. This still only gives us the ability to calculate the information a message carries about a state. Similar considerations apply to the calculation of the amount of information about an action. Ultimately, something close to knowledge of the whole game structure is needed to compute all the amount of information in a message. Thus, information is not a property of a message, as one may be misled to believe given the way the notion is presented; it is a property of the game. Speaking of messages or signals as *having* or *carrying* information can help us forget that.

This leads to another related issue that has to do with the availability of this information. From the modeler's perspective, who has a God's eye view of the signaling model, it may seem natural to think that "information is just *there*" (Skyrms, 2010, p. 44), but this is not necessarily the case for the agents involved in the game. This is duly noted by Skyrms:

> None of the probabilities used so far are degrees of belief of sender and receiver. They are objective probabilities, determined by nature and the evolutionary or learning process. Organisms (or organs) playing the role of sender and receiver need have no cognitive capacities. (2010, p. 44)

This is very important to keep in mind, since statements like saying that information is "just there" or saying that signals "contain" information can invite the idea that information is available to any kind of agent. As argued before, in order to calculate the measures of information proposed by Skyrms, one needs knowledge of the whole game. An agent that has less than perfect omniscient rationality does not have the same access to information as the modeler does.

There is one additional issue with the notion of information that needs to be pointed out before going back to more general considerations related to Wittgenstein's remarks. One should be careful in taking information as an indicator of successful communication, as is sometimes done in the literature (*e.g.* Wagner, 2012). In a simple Lewis signaling game, information about states and about actions is minimal for strategy profiles where sender and receiver uniformly randomize their choices, which is a scenario where intuitively one would say no communication is taking place. Conversely, information is maximal in a signaling system, a scenario of perfect communication. But information is also maximal in another situation. Consider again the examples in Figure 2.1. In the anti-signaling strategy profile (Figure 2.1b), the amount of information about states is maximal, since the sender strategy perfectly discriminates by using a different signal for each state. The same happens with the receiver strategy, so the amount of information about actions is also maximal. However, the anti-signaling profile is an example of an extreme case of miscommunication, where the receiver performs actions that are completely inadequate given the states.

This shows that one cannot blindly take high values of quantity of information about states and actions independently as direct proxies for successful communication between agents. The anti-signaling example is relevant because the metric is supposed to apply to any situation, not just strategies at equilibrium in fully cooperative scenarios. It is especially relevant when the argument is about information exchange in zero-sum games, where interests are completely misaligned, as is the case in the work of Wagner (2012). A better way of measuring successful communication is to calculate the amount of information between states and actions, as proposed by Godfrey-Smith and Martínez (2013). This requires one to merge the sender and receiver strategies into one function giving the probabilities that a given action is performed when a certain state obtains. An example of this merging operation is illustrated in Figure 5.1. Interestingly, Wagner's qualitative results still hold when using this measure instead (Martínez and Godfrey-Smith, 2016).

I believe that the notion of information is a useful one. However, because of the issues just raised, I think it is important to keep two things in mind when making use of it. First, that one should not speak of information content as something that signals have or carry. This invites misunderstandings of the same nature as the correspondence picture of meaning, and obscures the fact that a great deal of

$$\begin{tikzcd}
t_1 \arrow[r, ".21"] \arrow[dr] & m_a \arrow[r, ".46"] \arrow[dr] & a_1 \\
t_2 \arrow[r, ".63"'] \arrow[ur] & m_b \arrow[r, ".11"'] \arrow[ur] & a_2
\end{tikzcd}$$

(a) Mixed strategy profile.

$$\begin{tikzcd}
t_1 \arrow[r, ".80"] \arrow[dr] & a_1 \\
t_2 \arrow[r, ".27"'] \arrow[ur] & a_2
\end{tikzcd}$$

(b) Merged profile.

Figure 5.1: Example of how a mixed strategy profile for a simple binary signaling game can be merged. The probability that an action a is chosen when a certain state t obtains is calculated by taking into account the likelihood of each message m being chosen by the sender, and the likelihood of each of those messages leading the receiver to choose action a. Formally, $P(a|t) = \sum_{m \in M} \sigma(m|t) \times \rho(a|m)$.

knowledge about the whole particular signaling game is needed in order to calculate it. Second, one should be careful in taking information as an indicator of successful communication. I suggest that it is best to think of the amount of information about states (resp. actions) as a measure of the entropy, or conversely the degree of order or systematicity, exhibited by the sender (resp. receiver) strategy in the context of a game. Successful communication can be measured by the amount of information between states and actions, but this requires merging the sender and receiver strategies (see again Figure 5.1 for an example) and thus foregoing the idea that information is a property of signals.

With all that was said about information, it should be clear that I do not consider the notions of informational content or quantity of information as defining, or providing means to define, the meaning of signals in signaling games. Despite being notions that have a broader applicability than the notions of meaning provided by Lewis (1969) and Huttegger (2007b), and that can be useful in analyzing and understanding signaling game models and their dynamics, they are unable to capture every aspect of interest of what we intuitively would call meaning in a signaling game. This obviously precludes them from being a candidate for defining meaning in the broader context of natural language.

Not only do each of these proposals have specific problems of their own, they also share the more general issues stemming from attempting to conceive of meaning in terms of correspondence. It is surprising how strongly this picture keeps philosophers captive, and how often it keeps resurfacing, even in cases like that of Lewis, who originally saw the power of abandoning it and describing meaning in other terms, and Skyrms, who recognizes meaning as a “dangerous word” (2010, p. 34). Wittgenstein gives us at least three reasons why the correspondence picture is inadequate and we should abandon it. First, there is the issue of underdetermination in establishing relations between linguistic expressions and whatever their meaning is. Second, even if it would be possible to establish such a relation, that would not consist of everything that one needs in order to understand or make use of a linguistic expression in a successful way. Third, even if such a picture could successfully characterize certain types of use of linguistic expressions, in the broader heterogeneous and dynamic context of natural language, it would not cover all possible language games, and all the family of cases for which we use the word meaning.

In my opinion, the signaling games framework should never be taken as a theory of meaning, especially if the latter is seen as something yielding "a specification of the meaning of every expression and sentence" (Dummett, 1975, p. 1) of a language. The framework promotes the study of meaning not as a kind of correspondence, but rather in terms of use in a context with a driving purpose. One advantage of such a paradigm shift is that we are no longer tied to the need to explain how language hooks on to the world; this question is no longer relevant on this account. We can merely focus on trying to see how agents can use signals to cope with the world and achieve their purposes. On the way, we will better understand meaning, not by trying to say what meaning is, but by trying to better understand how communication works.

Chapter 6

Rules and meaning*

In this chapter, I address Wittgenstein's remarks regarding the role played by rules and rule-following in our mastery of language, and their implications for possible interpretations of signaling game models.

In order to understand Wittgenstein's remarks on rule-following, it is important to first try to clarify the conception(s) of rules that they are based on. As should be expected from the author's considerations on method (see again Chapters 1 and 3), no explicit definition is provided in the book. Furthermore, challenges put forward in various places against certain intuitions regarding rule-following are often implicitly anchored in different conceptions of rules. The discussion is, however, infused with analogies. A frequently used image is that of rules as tables, schemas, or charts, depicting a correspondence between linguistic and/or extra-linguistic elements (*e.g.* §1, §53, §62, §73, §86, §141, §§162-163, §265). Many remarks address the role of rules in actual games like chess (*e.g.* §31, §197, §200, §205, §§563-568) or tennis (*e.g.* §68). Another example of what is paradigmatically seen as a rule is a mathematical formula, like one specifying a number series (*e.g.* §§146-155, §§179-190, §226). The discussions around these examples usually address possible intuitions about the role rules play in the activities of which they are a part.

Wittgenstein's ultimate goal with these analogies is to illuminate a discussion on natural language. This is clear from how remarks directly about meaning and understanding are interwoven with considerations on rules in the book. Wittgenstein delves into the analogies not only because they illustrate general issues about rule-following, but also because he identifies them as representative of intuitions

*Wittgenstein's remarks on rules and rule-following in the *Philosophical Investigations* have motivated a prolific discussion in philosophy of language, at a certain point greatly fueled by Saul Kripke's influential, though highly contested, interpretation (1982). Despite its notoriety, I will here stay away from this debate and continue a mostly immanent reading following the choices laid out in Chapter 1. Good overviews of the debate are given by Miller and Wright (2002) and Hattiangadi (2007).

underlying some conceptions of meaning in philosophy. The rules of a game can be thought of as fully defining it, in the sense that they are sufficient and necessary knowledge for someone to be able to play the game; mathematical formulas can be imagined to characterize infinite cases and thus fully contain their own method of application as a self-sufficient mechanism; tables and schemas, like other rules, seem to force upon us a specific interpretation or behavior, leaving no room for choice; finally, rules can strike us as having an independent normative force that both underwrites our actions and establishes the distinctions between right and wrong, correct and incorrect. When it comes to language, similar intuitions are common: that the meaning of linguistic entities is fully defined, that knowledge of these definitions is a necessary condition to use or understand them successfully, that we have no choice in what our words mean, and especially in deciding whether an application or interpretation is correct or incorrect.

In the context of talking about language and meaning, Wittgenstein mentions a distinction that I believe can be useful in illuminating some of the general issues on rule-following that will be brought forward in this chapter. Despite not being given a sharp definition of rules, the following passage points to some possibilities:

> What do I call 'the rule according to which he proceeds'? – The hypothesis that satisfactorily describes his use of words, which we observe; or the rule which he looks up when he uses signs; or the one which he gives us in reply if we ask him what his rule is? (2009, §82, p. 43e)

This passage highlights some different perspectives on what we call following a rule. When someone engages in a behavior that seems to exhibit a regularity, we often say that they are following a rule. One way of being motivated to say that is if we have a hypothesis for describing the observed behavior, and perhaps even attempting to predict future behavior. For example, if I observe someone producing the sequence of numbers '$0, 2, 4, 6, 8, 10, \ldots$', I can advance the hypothesis that they are following the rule '+2', whereby one starts with the number 0 and subsequently adds '2' to the last number in the sequence in order to produce the next. This I will call the *descriptivist* perspective on rule-following.

We also say that someone is following a rule if they are using what we recognize as a specification of a rule as an instrument to guide their behavior. Imagine that the person producing the aforementioned sequence was at every step consulting a piece of paper with instructions in order to produce the next number in the series. This I will call the *material* perspective on rule-following. Note that, and this is a crucial point I will come back to later, these two possible cases do not necessarily overlap. I can advance a rule as a hypothesis to describe someone's behavior, and even if that description fits that does not imply that they must somehow, consciously or unconsciously, be making use of that very rule, or any rule for that matter, in a material sense. Conversely, if I observe someone apparently consulting instructions to produce some behavior, I may not necessarily be able to describe it or even see a regularity.

If we are dealing with communicative agents, we can have an additional motivation to characterize someone as following a rule. If I observe someone producing a certain behavior, I can simply ask them how they are doing it, and they can advance an explanation in terms of rule-following. I observe someone producing the sequence '$0, 2, 4, 6, 8, 10, \ldots$', and when I ask them how they are calculating the next number at each step, they tell me they are simply following the rule '$+2$'. Note that it is again possible that their reply does not align with either the rule we describe them as following, or with the rule they are actually using in a material sense. This *justificatory* perspective is different from the other two in the sense that it is first-person, rather than third-person. However, it can also be given in the descriptivist or material sense. One can characterize one's behavior in terms of a rule when one is making use of a specification of it as an instrument to guide one's behavior, or it can simply be provided as an *a posteriori* hypothesis.

In this chapter, I will discuss a bit more in depth Wittgenstein's remarks on rule-following focusing mostly on two parts of the *Philosophical Investigations.* Sections §§82-88 can be seen as a summary of some of the issues with common intuitions behind rule-following and some of Wittgenstein's proposals to clarify associated misunderstandings. Most of these are again discussed in more length in §§139-242, although remarks on, or related to, rules and rule-following can also be found in other parts of the book. In Section 6.1, I first discuss how most of Wittgenstein's remarks attempt to deflate various intuitions on the topic. Subsequently, in Section 6.2, I will try to characterize his more positive remarks on the role rules play in our practices. Finally, in Section 6.3, I reflect on the implications of these considerations for the signaling games framework.

6.1 Misunderstanding rules

One of the first mentions of rules in the *Philosophical Investigations* occurs in §31 and uses the example of chess. We are led to imagine showing someone a chess piece and saying "This is the king." This explanation, Wittgenstein observes, can serve to teach the other person how to use the piece but only if, for example, they already know the rules of chess and are only missing the knowledge of which particular piece in that context stands for the king. In this case, rules appear to have a foundational role: learning how to use the chess piece presented is anchored in the person's knowledge of the rules of chess. But what is to know the rules of chess? It is tempting to think that one has learned explicit instructions like "This is the king; it can move in this-and-this way" (2009, §31c, p. 19e). However, it is also possible to imagine the other person having learned to play chess empirically, by observing and imitating others, "without ever learning or formulating rules". In this case, the explanation "This is the king" would work just as well as in the former. The ostensive explanation thus teaches the use of the piece to the other person if they already know how to play the game, but knowledge of explicit rules is not necessary for that.

Similar remarks are made with respect to other examples of rules. In §143 we are introduced to a language-game involving two agents, A and B, where "when A gives an order, B has to write down series of signs according to a certain formation rule." (2009, §143a, p. 62e) Suppose A's order is to write down the series of the natural numbers in the decimal system, and we are testing B by observing their behavior. When can we say that B has understood this order correctly? Since the series is infinite, it seems that there is no one number that we can stipulate as a criterion such that, if B continues the series correctly up to that number, we can say that they have definitely mastered the series (§145). Is understanding then necessarily anchored in explicit knowledge of the algebraic formula? The following remarks (§§146-155) cast doubt on that idea. In particular, §151 advances again the suggestion that we can easily imagine someone continuing a series without the formula necessarily occurring to them. Think about whether you would need to call to mind a formula to continue, for example, the series of even natural numbers.

Related considerations are made about tables and charts. Consider the builder's language of §2 and the toy language-game of §48. One can imagine agents playing these games with the help of a chart associating elements of the game and signs (see §86 and §53). However, both games were originally imagined without appealing to such rules, and we plausibly had no need to imagine them when we were presented with the characterization of each game. It is additionally plausible to picture agents learning and using signs in the context of these games without the need for such tables. Although it is possible to introduce them, it is not always necessary. One can understand and play either of these language-games without the need for explicit rules.

Learning and playing chess, reproducing a number series, associating objects or colors with words, those are all activities that involve repeatable patterns and regularities. One of the upshots of Wittgenstein's considerations is a plain reminder that it is perfectly imaginable that any of these activities can be mastered without the need for explicit rules to anchor our understanding of them (see §§208-209). It serves as a first palliative to the urge to hypostatize rules from regularities and conflate the various conceptions of rule-following. If one can describe some behavior in terms of rules (descriptivist conception), this does not imply that a rule is necessarily anchoring that behavior (material conception). Similarly, just because someone justifies their own behavior by appeal to a rule (justificatory conception), that does not imply that they were making use of that rule when performing the behavior (material conception). These considerations can be summarized in the following remark:

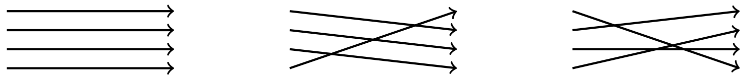
Remark 6.1.1 *Producing behavior that exhibits regularities does not necessarily require knowledge of explicit rules.*

Even if one agrees that knowledge of explicit rules is not necessary for understanding regularities or behaving accordingly (playing chess, expanding a number series, and so forth), it is possible to still hold that it is nevertheless sufficient. When one has knowledge of an explicit rule, our behavior when following it can easily appear to be fully determined and guided solely by the rule. Against this idea, Wittgenstein repeatedly raises various incarnations of an objection that is aptly summarized in the following passage:

> This was our paradox: no course of action could be determined by a rule, because every course of action can be brought into accord with the rule. The answer was: if every course of action can be brought into accord with the rule, then it can also be brought into conflict with it. And so there would be neither accord nor conflict here. (2009, §201a, p. 87e)

One of the first illustrations of this point can be found in §86, where a variant of the builder's language of §2 is introduced. In this toy language-game, the builder A shows the assistant B written signs, and B makes use of a table to handle them. This table has two columns: one with signs, and another with pictures of building stones. B uses this table by looking up the signs received from A in the first column, tracing the opposite picture in the second column, and handing A a building stone that looks like that picture. The table works as a rule: it appears *prima facie* to fully determine how B should interpret the written signs given by A.

Is the table sufficient to condition B's interpretation of the written signs? One presumes the elements in the two columns are to be related horizontally, as illustrated by the schema in Figure 6.1a. However, it would be possible to relate them



(a) Horizontal (b) Alternative (c) Random

Figure 6.1: Interpretation schemas for the builder's language variant of §86.

according to an alternative, but apparently systematic schema, for example the one illustrated in Figure 6.1b, or even according to an apparently random schema like the one illustrated in Figure 6.1c. So, it now seems that the table is not sufficient to determine how to interpret the signs, since it is itself subject to different interpretations. What about the table together with the schema, could that alone determine the interpretation of the signs in the language-game? One needs to realize that the schema is itself also made of signs, which in turn could be interpreted alternatively. Do we also require an additional rule to interpret the schema? We are on the path to an infinite regress.

The possibility of variously interpreting a table is briefly raised again in the context of the example of copying text from print to handwriting (§§162-163) and discussed a bit more extensively in relation to algebraic formulas, namely in the case of the example of learning to write down a number series (broached in §143 and picked up again in §§185-186). In particular, we are invited to consider the following case. Say a tutor A is teaching a pupil B to write down series of the form '$+n$', such that A expects B to write down '$0, 1, 2, 3, 4, 5, \ldots$' when given the order '+1', '$0, 2, 4, 6, 8, 10, \ldots$' when given the order '+2', and so forth. Imagine that B has demonstrated to have correctly expanded different series up to the number 1000, but when tested again for '+2' beyond 1000 he writes '1000, 1004, 1008, 1012, . . . '. About this situation, Wittgenstein writes:

> [. . .] we might perhaps say: this person finds it natural, once given our explanations, to understand our order as *we* would understand the order "Add 2 up to 1000, 4 up to 2000, 6 up to 3000, and so on". (2009, §185c, p. 81e)

Even though *we* would probably continue the series by writing down '1000, 1002, 1004, 1006, . . . ', there is no principled reason why B's alternative interpretation is less valid than ours. The point here is that the algebraic formula '+2' apparently does not, in the thought experiment, determine its own interpretation by itself; it was subject to different interpretations by A and B. Although A's interpretation feels more correct to us, Wittgenstein suggests this is just a matter of familiarity. B's interpretation feels like if someone would react to a pointing gesture "by looking in the direction from fingertip to wrist, rather than from wrist to fingertip." (2009, §185d, p. 81e) However, there is in principle no guarantee that either order could not be interpreted otherwise. If one is tempted to postulate a further meta-rule to constrain those interpretations, we again face the possibility of an infinite regress.

These issues naturally carry over to language and meaning. Explicit connections are made in various places. We find a clear statement on that in the following passage:

> Speaking of the application of a word, I said that it is not everywhere bounded by rules. But what does a game look like that is everywhere bounded by rules? whose rules never let a doubt creep in, but stop up all the gaps where it might? – Can't we imagine a rule regulating the application of a rule; and a doubt which *it* removes – And so on? (2009, §84a, p. 44e)

These remarks connect the intuition that a game is completely delimited by its rules, with the expectation that a language-game similarly fully constrains the ways words and sentences can be used within it (see again Chapter 4). This relates to our current discussion because a language-game that would be everywhere bounded by rules, that would never let a doubt creep in, would be a game that fully determines the application of words, and hence their meaning, within that game. But, we are again confronted with the suggestion that we can always entertain a doubt, if not in the application of a word, maybe in the application of a rule that supposedly regulates the application of the word, and so forth.

Do all of these problems throw doubt on Wittgenstein's own notion of language-game and his propounded connection between meaning and use? Other passages show that problems only arise if one is tempted to identify meaning with use or misunderstand the role of use in language production and understanding. In §§139-141 we are presented with questions about whether we grasp the whole use of a word like 'cube' when we understand it, and whether use can therefore determine meaning. If one imagines that a picture of a cube comes to mind, does that picture leave no doubt as to which situations would warrant the use of the word 'cube'? Wittgenstein suggests that the picture of a cube could fit a triangular prism under a certain method of projection. Should the whole use then consist of picture plus method of projection? In §141 we are again confronted with the idea that this purported solution simply triggers an infinite regress. Although we might have an intuition that we grasp the whole use of a word at a stroke, Wittgenstein defends that we do not have a concrete model of how such a thing would actually be possible (§§191-192). A word cannot fully and unequivocally determine its own meaning, just as no explicit rule is sufficient to fully determine the actions that are carried out in response to it.

A foundationalist reply to these considerations would suggest that such problems simply show that some rules must, therefore, be fundamental; there needs to be a backstop to the infinite regress. This idea is addressed by Wittgenstein by discussing the possibility of a rule, or an interpretation thereof, containing its own method of application. This would supposedly stop the potential infinite regress by making the rule self-sufficient. Going back to the algebraic formula '+2', if the tutor corrects the pupil when he writes '1000, 1004, 1008, 1012, ...', doesn't that show that he already knew in advance what values to expect? In §§186-190, there is an attempt to flesh out this intuition. Is it implied that the tutor knew all the possible values in the series? This can hardly be the case, since the series is infinite. Did the tutor know a procedure that dictates how to obtain any number in the series from its predecessor? But that is exactly what the explicit rule was supposed to do: to determine what follows from one stage to the next. If knowing how to use the rule requires knowing how to follow this procedure, how do we then know how to interpret the procedure? There doesn't seem to be a clear picture of what it would be like for a rule to contain its own method of application without getting us back to the problem of interpretation.

In the following paragraphs, Wittgenstein makes another analogy:

> A machine as a symbol of its mode of operation. The machine, I might say for a start, seems already to contain its own mode of operation. What does that mean? – If we know the machine, everything else – that is the movements it will make – seem to be already completely determined. (2009, §193a, p 83e)

The picture of a machine illustrates a self-contained rule: it seems that its mechanism is built in, that its parts can only move in one way, that its future behavior is fully pre-determined. But we need only remember the possibility of the parts "bending, breaking off, melting, and so on" (2009, §193b, p 84e) to realize that this is not necessarily so. In these cases, there is a possibility that the machine behaves in a way that is different from what we would expect, either from the way it is built, or from our familiarity with the mechanism. Similarly, even though the beginning of a series might strike us as "a visible section of rails invisibly laid to infinity" (2009, §218, p. 91e), we should not forget that it is always possible that at each stage the rails bend, break, or melt, or that it strikes others as a different track altogether.

These insights can be generalized in the following remark[2]:

Remark 6.1.2 *A rule can be variously interpreted in any case.*

These considerations focus on an aspect of rule-following that relate to whether or not it is possible for a rule to be interpreted in more than one way. Rules are also typically perceived by us as having a normative side. Even if it is possible for the aforementioned pupil to interpret '+2' differently than we would expect, there is an additional feeling that they would be wrong in doing so. This is closely intertwined with the previous remark, but raises further issues. Consider the example of teaching a series. In §143 we are asked to imagine how this would go about. The series of the natural numbers in the decimal system could be taught by requiring the pupil to copy numbers previously written down for him. In this learning process, it is always possible that the pupil makes mistakes. When given the first 10 digits as '$0, 1, 2, 3, 4, 5, 6, 7, 8, 9$', examples of mistakes would be to write down '$8, 2, 9, 4, 5, 1, 3, 7, 6, 0$' or '$1, 0, 3, 2, 5, 4, 7, 6, 9, 8$'. Although there does not appear to be any discernible order in which the numbers were supposedly copied in the first example, the second mistake could be attributed to the pupil swapping the numbers pairwise. The latter behavior could be described as rule-following, but one would also like to say that the pupil in that case, although following a rule, is following the wrong rule.

One first problem has to do with the binary mindset normative issues usually motivate. In the context of the aforementioned example, it is tempting to separate possible mistakes into random and systematic. Wittgenstein invites us to notice that, however, "there is no sharp distinction between a random and a systematic mistake" (2009, §143c, p. 62e). If the pupil repeatedly produces the sequence '$1, 0, 3, 2, 5, 4, 7, 6, 9, 8$' we would like to say they are making a systematic mistake. But what if they produce that sequence once and then repeatedly produce '$0, 1, 2, 3, 4, 5, 6, 7, 8, 9$'? Should we call the first event a random mistake? What if they do it every 1000 tries? How often does the pupil need to do it as we expect for us to be completely sure they have mastered the system? And even if the sequence is reproduced to our satisfaction once, how can we rule out the possibility that this was done by chance?

These remarks are reiterated in the context of other examples. In §157, we are asked to imagine the process of training pupils to read mechanically, *i.e.* producing sounds from written text. Here too, through the different stages of learning there is no clear distinction between a pupil that produces random sounds, one that produces

[2] Adapted from a passage in §28, and following Stern (2004, p. 142).

the expected sounds but only by accident, and one that is already able to read. In §163, in the case of using a table to copy from printed text to handwriting, again we are confronted with cases which we would probably classify as simple systematic mistakes, like using the table according to the schema in Figure 6.1b rather than the one in Figure 6.1a. One could imagine additional variations further complicating the systematic mistake, potentially to a point where one would be unable to demarcate systematic from random[3].

The effect of these remarks is to undermine the binary character that our normative talk about rule-following seems to impose on experience. Although it seems that whether someone is following a rule or not should be subject to a this binary classification, a sharp distinction between behaving randomly and following a rule correctly is difficult to pin down precisely. Because of that, we might feel the urge to project the distinction onto the rules themselves. Maybe we cannot always say whether an application of a rule is correct or incorrect, but surely, we feel, there are criteria of correctness corresponding to the rule. However, if we just consider Remark 6.1.2 (p. 149) again, we see how it is problematic to think that rules come with their own criterion of correctness. If the rule cannot by itself determine its own interpretation, it certainly cannot determine which possible interpretations are right or wrong.

Similar considerations are made in §§139-141 about the aforementioned example of the cube. In particular, the question of whether or not a mental picture of a cube could determine the use of the word 'cube' is also a question of normativity. The example of a method of projection under which one could consider the use of the word 'cube' when pointing to a triangular prism is advanced to defy the intuition that, when one understands the word 'cube' (in this case as a picture in the mind) one is in possession of the criteria establishing the correct uses of the word:

> What was the effect of my argument? It called our attention to (reminded us of) the fact that there are other processes, besides the one we originally thought of, which we should sometimes be prepared to call "applying the picture of a cube". So our 'belief that the picture forced a particular application upon us' consisted in the fact that only the one case and no other occurred to us. (2009, §140b, p. 61e)

[3] Case in point, the sequence '$8, 2, 9, 4, 5, 1, 3, 7, 6, 0$' was created using a pseudorandom number generator, a computer program which follows a deterministic algorithm to produce, from a so-called seed value, sequences of numbers that share properties with truly random sequences. A pupil producing such a sequence could thus ultimately be interpreted as making a systematic mistake. One would need only to find the right value for the seed that would generate such a sequence, and describe the pupil as following a rule: the pseudorandom algorithm.

The belief that language contains in itself a normative force is to a certain extent attributable to the feeling we are struck with when encountering familiar words. But an exercise in imagination can make us see that there are additional cases that, although at first sight seem to be incorrect uses (*e.g.* 'cube' for a triangular prism), can be made correct without changing our understanding of the original word (*e.g.* by considering different methods of projection). These observations can be summarized in the following remark:

Remark 6.1.3 *A rule cannot contain its own criteria of correctness.*

In order to strengthen these remarks, Wittgenstein accompanies them with a diagnosis of the origins of the problems. Unsurprisingly, the intuitions criticized are seen as arising from the same source as philosophical confusions in general (see again Section 3.1): misunderstanding language and the ways we talk. This is first broached in §81 in relation to logic. Wittgenstein observes that "in philosophy we often *compare* the use of words with games, calculi with fixed rules" (2009, §81a, pp. 42e-43e). This, he claims, can mislead one into thinking that language approximates such calculi. An example is the expectation that formal logic could provide an accurate model of natural language. But, Wittgenstein insists:

> [...] logic does not treat of language – or of thought – in the sense in which a natural science treats of a natural phenomenon, and the most that can be said is that we *construct* ideal languages. (2009, §81a, p. 43e)

Overlooking the status of the comparison and taking the analogy too far can lead us into "thinking that if anyone utters a sentence and *means* or *understands* it, he is thereby operating a calculus according to definite rules" (2009, §81b, p. 43e). Even if formal logic could be used to describe regularities in natural language in terms of rules, this would not imply that someone who is using natural language is following those rules in a material sense. The latter is not necessarily the case, as the previous remarks emphasize.

We are again facing the issue of conflating different perspectives on rules. To describe someone's behavior as following a rule is simply to provide a hypothesis or a model to sum up what one observes (§156g). We do this by identifying regularities and paying attention to characteristic signs of mistakes and correct actions in how others react to them (§54), which we identify based on our knowledge of shared human behavior (§206c). But these kind of descriptions need not (and should not) be taken as committed to the idea that the agents so described are making use of rules in a material sense, nor that they would necessarily justify their behavior in terms of those rules that we attribute to them.

Another source of misunderstandings stems from our familiarity with natural language. One reason why we might feel that a rule compels a certain application, or that a word determines its own usage, is our being used to applying the rule or using the word in a certain way, which can blind us to the possibility of alternatives. The case of the application of the word 'cube' (§§141-142) was already discussed above. Another clear statement of this issue pertaining to the case of reading can be found in §§166-171, where Wittgenstein contrasts the cases of reading a familiar versus a strange sign. Whereas an unfamiliar sign does not immediately elicit a specific sound to us, the feeling is different with a letter of the alphabet we normally use or words in a language that we do recognize. In the latter case, it feels like the connection is not merely arbitrary. Familiar signs make a *deep* impression on us (§167), we have the feeling that they *cause* our utterances (§169), as if the signs themselves *guide* our behavior (§170) and *intimate* their sounds to us (§171).

These are possible ways one would describe the experience of reading a familiar sign when trying to look closer into the activity of reading. But we need to be careful not to be misled into attributing these properties to our actual experience:

> But now, just read a few sentences in print as you usually do when you are not thinking about the concept of reading, and ask yourself whether you had such experiences of unity, of being influenced, and so on, as you read. – Don't say you had them unconsciously! Nor should we be misled by the picture of these phenomena coming forth 'on closer inspection'. (2009, §171b, p. 76e)

With the association sign-sound serving as an example of a rule, the danger seems to lie in conflating the justificatory with the material conceptions, *i.e.* in taking our *a posteriori* justifications of rule-following behavior as evidence that our actions made explicit use of rules all along.

Another example reiterates this point in §§175-178. Wittgenstein asks us to make an arbitrary doodle on a piece of paper, and subsequently make a copy of it by letting ourselves be guided by the original. Again, it is said that nothing special seems to be going on while we are performing this task. But afterwards, when we reflect upon it and are asked (or ask ourselves) how it was that we managed to do it, we are tempted to explain the experience in certain terms, and project elements of that explanation back onto the experience. This is not because we necessarily remember it as such, but because we are looking at it through the medium of the terms that we use to provide such explanations (§177).

Finally, this issue is mentioned again in the context of the example of a machine:

> When does one have the thought that a machine already contains its possible movements in some mysterious way? – Well, when one is doing philosophy. And what lures us into thinking that? The kind of way in which we talk about the machine. (2009, §194a, p. 84e)

The point is clear, and applies to the problems of rule-following in general. The issues raised in Remarks 6.1.1, 6.1.2 (p. 149), and 6.1.3 (p. 151) attempt to dismantle two intuitions: first, that knowledge of a rule is necessary and sufficient for generating or understanding behavior that exhibits a regularity; and second, that a rule can be self-sufficient and carry with it its own criteria of correctness. According to Wittgenstein, these intuitions stem from the way we describe regularities or justify or own behavior in terms of rules. Conflating these descriptive and justificatory practices with a material conception can mislead us into hypostatizing rules as necessarily underlying regularities in our practices, and attributing them inflated powers in those practices:

Remark 6.1.4 *We are misled into hypostatizing rules and projecting special powers onto them by the ways we talk about behavior that exhibits regularities.*

In summary, Wittgenstein's perspective seems to be that rules are often misunderstood. Some of our behavior exhibits regularities. We often talk about this type of behavior in terms of rules: we can describe others, or justify our own actions, as following a rule. To take these descriptions or justifications as evidence that rules necessarily underlie such behavior is to incur in a misunderstanding. Once we make that leap, another confusion that can easily follow is to think that rules have the ability, by themselves, to guide or even force our choices and actions. These misguided intuitions can form part of a picture of language where linguistic entities have the capability independently determining their own meaning. In order to resist these misunderstandings, Wittgenstein intersperses the discussion with additional remarks that give hints towards an alternative conception of rules and rule-following.

6.2 An alternative picture

The path to an alternative picture of rule-following can start by reminding ourselves that there are various roles rules can play in different language-games, and that there are various activities we can characterize in terms of following a rule. For example, in reference to the language-game of colored square of §48 (see again Section 3.2), in §53 Wittgenstein imagines different scenarios in which one could say that a sign in that game names a square of a certain color. This includes the teaching of the game being done in a particular way, or there being a chart establishing a correspondence between signs and elements in the game, to be either used in teaching, resolving disputes, or as an actual tool in the use of the language. This is subsequently reiterated for rules in general:

> Just think of the kinds of case where we say that a game is played according to a particular rule.
> The rule may be an aid in teaching the game. The learner is told it and given practice in applying it. – Or it is a tool of the game itself. – Or a rule is employed neither in the teaching nor in the game itself; nor is it set down in a list of rules. One learns the game by watching how others play it. But we say that it is played according to such-and-such rules because an observer can read these rules off from the way the game is played – like a natural law governing the play. (2009, §54, p. 31e)

These scenarios are examples of cases where we are likely to characterize practices in terms of rule-following according to the material and descriptive conceptions of rules. We can imagine an explicit rule as a tool, either involved in the teaching of a practice, or used in the practice itself. But we can also imagine both activities conducted without any employment of an explicit rule, but in a way that an external observer could describe as involving rules by extrapolating those from the observed regularities. Wittgenstein makes this observation, I believe, as a reminder for us not to conflate the two.

We are also confronted with various possibilities regarding other examples of activities that we might characterize as involving rule-following. In §151, we are given several descriptions of the behavior, and hypotheses of the mental experiences, of an agent B that is trying to find a rule for number series used by another agent A. As A writes down series of numbers, B can mentally try out explicit algebraic formulas until the next number confirms his suppositions, he can be watching "with a certain feeling of tension" while "all sorts of vague thoughts float through his head" (2009, §151c, p. 65e), or he simply recognizes the series with ease without even thinking about it. Similar scenarios are presented in §§156-171 for the activity of reading, and in §172 for the experience of being guided.

The exploration of the various experiences one might go through when involved in an activity that could be characterized as following a rule, and of the alternative ways an external observer might describe such experiences, serves to suggest that we might be in the presence of a family of cases (see in particular §164) and that, therefore, there might not be one feature that occurs in all cases. The lesson is that "we must be on our guard against thinking that there is some *totality* of conditions corresponding to the nature of each case" (2009, §183, p. 80e). It is thus important to keep the following in mind:

Remark 6.2.1 *Rules can have various roles in a game; there is a diverse number of activities and experiences that we can characterize as following a rule.*

Another observation that can help us form a different picture of rule-following is realizing that when we follow a rule, even an explicit one, we ultimately do so blindly (§219). One might feel a certain discomfort in following the regress arguments advanced by some remarks in the previous section. This is because, in our everyday practices, when we make use of a table in a way similar to the language-game of §86, we don't feel the need for an infinite chain of additional schemata to clarify how to interpret the table; we simply use it. When we need to continue a series that we are familiar with, we act with perfect assurance, without being troubled by the lack of self-sufficient reasons (§212). Even if there are steps of interpretation, we eventually reach a point where we simply need to stop the regress in order to act. In general, at a certain step in the chain, we ultimately "read the lips of the rule and *act*, without appealing to anything else for guidance" (2009, §228, p. 93e).

Additionally, when we justify our behavior in terms of rules, we are also not prepared to further appeal to rules for those rules, and so on *ad infinitum*. If asked why we produce certain sounds when reading text in a familiar language, we justify it by the letters which are there (§169b) and probably cannot give any further reasons. Once we exhaust our justifications, and we eventually do, we are left with nothing more than the observation: "This is simply what I do." (2009, §217, p. 91e) This is not to deny that we can sometimes characterize one rule in terms of another, but simply to observe that even when we do have a chain of such justifications, at some point we are at a loss to provide more.

How do we get to produce rule-like behavior, or get to understand regularities in terms of rules as we do? Wittgenstein's answer is that we learn this simply by training. The pupil in the language-game in §86 learns to use the chart by being trained, which may include for example learning to pass his finger horizontally from left to right. We also learn to use a number series by a process of training like the one described in §145 (see also §189b). Thus a causal connection between explicit rules and rule-following is established: we are trained to react in a certain way to a particular rule or sign, and so we do react to it (§198). Wittgenstein's description of the process of teaching a rule to a pupil is as mundane as it is intuitive:

> I do it, he does it after me; and I influence him by expressions of agreement, rejection, expectation, encouragement. I let him go his way, or

> hold him back; and so on.
>
> Imagine witnessing such teaching. None of the words would be explained by means of itself; there would be no logical circle. (2009, §208cd, p. 89e)

We can all recognize such a process by having been on at least one side of it in our lives. And we can then see how it is possible to learn to follow a rule without the need for an infinite chain of explanations or interpretations; the process is simply anchored in shared human behavior (§206).

An important thing to note about this perspective is that, although a backstop for the regress problem is being suggested, it is of a different kind of that proposed by the hypothetical foundationalist. It is not that certain rules are special and thus immune to reinterpretation. The point summarized in Remark 6.1.2 (p. 149) still holds. It is rather that when an agent follows a rule they have, at some point in the infinite chain of possible re-interpretations, acted without questioning an interpretation. A possibility of doubt need not force one to doubt. One can take the chart of §86, appreciate the various possible schemas of interpretation, but simply choose one and use it to determine which building stone to hand over when observing a certain sign. Naturally, because of the possibility of other interpretations, success is not fully guaranteed. But, given one's training and experience in the practice, it is possible to have a feeling of how likely one is to get it right. It is also a highly contingent and subjective backstop. Just because one person in one instance decided to stop the regress there, does not imply that they will do so in other circumstances, or that others do exactly the same.

We can summarize these points in the following remark:

Remark 6.2.2 *When we follow a rule, we ultimately do so blindly, and learning to do so is a matter of training.*

But there is more to Wittgenstein's positive observations. There a number of characterizations of rule-following in terms of regularities (*e.g.* §207, §237), customs (§198, §199, §205, §337), and practices (§197, §202). Starting with the first, in §237, we are asked to imagine someone intently following a line in a way that reveals no regularity to us. About this scenario, Wittgenstein says:

> We can't learn his way of following the line from him. Here perhaps we really would say: "The original seems to *intimate* to him how he has to go. But it is not a rule." (2009, §237, p. 94e)

Note that this is a remark related to the descriptive conception of rules. From a third-person perspective, one describes some behavior as conducted according to a rule only if one can recognize a regularity. The agent might justify his own behavior in terms of following the line according to a certain rule. And even if he doesn't, it is conceivable that he could be making use of an explicit rule as an instrument of the particular game he is playing. But if we are unable to observe any regularity in his behavior, we are unlikely to describe him as following a rule. This is perhaps the reason why, for Wittgenstein, the behaviors we call following a rule do not occur on only one isolated occasion (§199): one cannot identify a regularity based on a single occurrence.

These observations extend to language as well. In §142, it is pointed out that our language-games rely heavily on regularities. In particular, regularities allow us to expect certain words to be used in a certain way, and to predict a particular use of a word to be likely to be successful. Wittgenstein talks about normal and abnormal cases. The former are those situations where the use of a word seems clearly laid out in advance. We are used to observing others uttering a certain word in a certain context, or have past experience in using or reacting to a word in a certain way, and expect these regularities to hold up again in the future. In abnormal cases, doubt creeps in. But the way we use language in normal cases is not necessarily invalidated by the abnormal cases. As mentioned before, just because a doubt is possible does not imply that we need to constantly be in doubt. Our normal use of language still holds as long as there are regularities to rely on. If that was not the case, however, our language-games would, according to Wittgenstein, lose their point (§142). Being able to identify regularities is also crucial for calling any observed practice a language (§207).

What we call following a rule involves not only a regularity for a single individual, it involves other individuals as well. In this regard, §199 is key:

> Is what we call "following a rule" something that it would be possible for only *one* person, only *once* in a lifetime, to do? – And this is, of course, a gloss on the *grammar* of the expression "to follow a rule".
> It is not possible that there should have been only one occasion on which only one person followed a rule. (2009, §199b, p. 87e)

The purpose of these remarks is to remind us that there is more to rule-following than an explicit rule and an isolated individual purportedly following it. For a behavior to be characterized in terms of following a rule, it is not only important that it is anchored in and recognized as a regularity, but also that this regularity is a custom, *i.e.* an established usage between more than one person. One characterizes some behavior as following a rule only when it is taken in this broader context. A custom, like a practice, has a history behind it, and it is passed on and upheld by a group. To say that someone is following a rule is to say that that individual is participating in that practice (§202). Note that these statements are not intended as metaphysical claims. They are rather, as is pointed out in the same passage, a remark on how the expression 'following a rule' is normally used.

As this applies to practices that involve rule-following in general, it also applies to language in particular:

> To follow a rule, to make a report, to give an order, to play a game of chess, are *customs* (usages, institutions).
> To understand a sentence means to understand a language. To understand a language means to have mastered a technique. (2009, §199bc, p. 87e)

Making a report and giving an order are part of Wittgenstein's list of examples of language-games (§23). The rules of language ultimately rest on this edifice of convention as well (§355). And this is crucial to realize in order to understand meaning. Words and sentences, to be used meaningfully, need to be embedded in the larger context of a language (see also §337). This context is usually implicit in the ways we behave, it forms the scaffolding that makes communication possible (§240). We can summarize these general points about rules in the following remark:

Remark 6.2.3 *Using a rule is a custom, a practice backed by regularities and an established usage within a group.*

It may seem that Remark 6.2.1 (p. 155) precludes the advancement of these general positive characterizations of rule-following. If all the practices we characterize as following a rule belong to a family of cases that do not necessarily share one common set of properties, how can one justify the general claims of Remarks 6.2.2 (p. 156) and 6.2.3 (p. 158)? The relevance of the first remark is exactly to keep one from taking the remaining observations as forming the basis of a theory of rule-following. They simply serve as observations to keep in mind when reflecting on rule-following behavior to avoid falling prey to the urges of precision and universality that can lead to the misunderstandings identified in the first place.

6.3 Rules in signaling games

How are these considerations relevant in the context of the signaling games framework? We can start by thinking of how signaling games could potentially illuminate Wittgenstein's remarks. In his discussion of rule-following, Wittgenstein often uses tables, schemas, or charts as examples of rules. One can see these as devices that make explicit some direct relation between two sets of elements, like signs and building stones in the game of §86. One immediate parallel that comes to mind when considering this kind of examples is between rules and signaling strategies. Much like these examples of rules, signaling strategies establish relations, either between states and messages (for the sender), or between messages and actions (for the receiver). But, as we have seen in Chapter 2, there are many possible ways of interpreting what these strategies represent.

Lewis (1969) based his considerations on the traditional game-theoretic approach of speculating about individual agents making rational choices in an interactive setting. Within this picture, he viewed signaling strategies as possible contingency plans. Agents would consider alternative strategies, calculate their expected utilities, compare them to each other (for example in the manner illustrated in Table 2.1), and choose one to commit to for playing the game. The chosen strategy presumably serves, during actual play, as a guide that the sender (resp. receiver) makes use of for deciding the choice of message (resp. action) when confronted with a certain state (resp. message). This evokes a lot of similarities with the material conception of rules. Strategies can be seen as instruments, explicit rules used by agents to guide their behavior.

There are many practices where we do make use of and have to coordinate on explicit rules. Suppose Alice and Bob, from the example given in Section 2.1, cannot communicate directly since they are meeting in secret. Bob can see Alice's balcony from his window, so they sit down to agree on a code Alice can use to signal Bob in the future whether she wants to go to Café One or Bistro Two. Alice can hang either a red or a blue scarf on her balcony, and Bob will use this to decide where he goes to meet her. They analyze their options by considering something like Table 2.1, conclude that they are better off using one of the two possible signaling systems. They choose one, each writes their own part down, and go their separate ways. The next day, Alice wants to meet Bob in Café One. Not remembering the agreed upon code, she consults the piece of paper where she wrote it down, and hangs a scarf of the appropriate color on her balcony. Bob, upon seeing the signal (and also having a bad memory) looks up in his own piece of paper which place he is supposed to go to given the color of the hung scarf.

In this example, Alice and Bob are each following rules in a material sense. Alice wrote down a sender strategy, and Bob a receiver strategy. The scenario can be modeled as the binary signaling game presented in Section 2.1. This in turn can be used to make explicit why Alice and Bob behave the way that they do, or to speculate on a number of other issues. One can use to model to explore, for example what could happen if one or both were to lose their papers, what the role of repeated play and memory limitations could be, what could happen if preferences suddenly changed, among many other things. These are all valid and potentially interesting questions to explore. Problems only arise if one tries to overgeneralizes this picture. As Wittgenstein reminds us, not all instances of what we call following a rule require knowledge or use of explicit instructions (Remark 6.1.1, p. 145). This would make such a picture limited to the cases where we do. Additionally, if this story is supposed to illuminate questions of meaning, one should not forget that explicit rules can be variously interpreted (Remark 6.1.2, p. 149). The material picture does not give us a better understanding of how the strategy itself is understood by the agents.

I have been discussing the possibility of seeing strategies as rules, but this might be too narrow. Many characteristics of strategies that motivate the comparison depend on their place as a part of a broader context of a signaling game. Perhaps the appropriate comparison is rather between rule-following and conventions. This is something that Lewis himself explores (1969, pp. 100-107). He suggests that most conventions[4], including those involving signaling, can be naturally understood as rules. However, he also argues that not everything we call a rule can be framed in terms of convention, and gives a number of counterexamples. He concludes the following:

> We might be tempted to try distinguishing several senses of the word "rule," hoping that one of them would agree with my definition of convention. I doubt that the project would succeed. [...] We seem to be dealing with an especially messy cluster concept, and one in which the relative importance of different conditions varies with the subject matter, with the contrasts one wants to make, and with one's philosophical preconceptions. (1969, p. 105)

Prima facie, a comparison between rule-following and Lewis' notion of convention would be inadequate according to the author's own opinion. Sillari (2013) argues that, nevertheless, "all rules pertinent to Wittgensteinian rule-following involve a conventional element and hence can be analyzed as pertaining to situations [...] consistent with Lewis's analysis of convention in terms of coordination" (2013, p. 876). I think that an appreciation of Remark 6.2.1 (p. 155) shows that Wittgenstein probably had a similar position to Lewis' on this matter, and that any attempt at *equating* rules and conventions is likely to be unsuccessful. But, even if not *all* rules are con-

[4] Lewis mentions that there are potential exceptions to this (1969, pp. 104-105).

ventions, nor vice versa, I agree that the two notions are close enough to motivate a comparison, even if Lewis would disagree.

Consider Lewis' first definition[5] of convention:

> A regularity R in the behavior of members of a population P when they are agents in a recurrent situation S is a *convention* if and only if, in any instance of S among members of P,
>
> 1. everyone conforms to R;
> 2. everyone expects everyone else to conform to R;
> 3. everyone prefers to conform to R on condition that the others do, since S is a coordination problem and uniform conformity to R is a proper coordination equilibrium in S. (1969, p. 42)

There are a lot of connections between this definition and Remark 6.2.3 (p. 158). Wittgenstein's appeal to customs and practices is designed to draw attention to the idea that one can be said to follow a rule when one's behavior is embedded in a context of the kind spelled out by Lewis. To follow a rule is to repeat a pattern of established behavior in a population. The behavior is a known regularity as a response to a recurrent situation which, for that very reason, is expected by other agents in the population to be repeated. In a Lewis convention, conformity is preferred since it is the optimal behavior for all agents involved. In rule-following, this might not always be the case. Lewis's definition is obviously stricter and more precise than Wittgenstein's appeal to customs[6], but they both evoke similar ideas.

One aspect of rule-following that can be better understood by looking into research on signaling games has to do with the notion of blind action. One of Wittgenstein's remarks on this topic is a reminder that, when we follow a rule, our behavior is ultimately anchored in conduct that is not a process of interpretation (Remark 6.2.2, p. 156). Sillari (2013, pp. 885-888) argues that the evolutionary game theory approach to signaling games gives us a way to capture this insight[7]. Instead of thinking of strategies as instruments used by Alice and Bob, one can see them as simply representing regularities observed in the behavior of individual agents or populations thereof. This interpretation need not make any assumption about the cognitive ca-

[5] The final definition (1969, p. 78) is more refined, but for the purpose of this discussion, I believe that this first rough definition should suffice.

[6] See also the contrast with other related notions (agreement, social contracts, norms, conformative behavior, and imitation) given by Lewis (1969, pp. 83-121).

[7] Although I agree with this point, I disagree with the application to Kripke's paradox provided (Sillari, 2013, pp. 887-888).

pacities of those agents. We can see them as black-boxes, and ultimately even think of their behavior as simply hard-wired.

When we move from a material to a descriptivist stance in this way, the problematic step of interpretation is removed from the picture. Strategies simply capture what agents do, while the modeler is agnostic about the way they do it. As Skyrms (1996) and others have shown (see again Chapter 2), coordination of signaling strategies can emerge in a variety of situations even between extremely simple agents with hard-wired behavior. Signaling games can help illuminate how regularities in behavior between multiple agents can arise as a result of adaptive processes of various kinds. In particular, this can happen even when one assumes the behavior of the agents is blind. One may have qualms about calling strategies, under this interpretation, rules. But they do fit the descriptivist conception, as they are hypothesis that describe an agent's use of signals. In a signaling equilibrium, agents could be said to be following rules as their behavior exhibits regularities in the context of a practice. And as rule-following is a family of cases (Remark 6.2.1, p. 155), such an interpretation could be said to capture some of those cases. Strategies, under this interpretation, can at least be said to capture behavior that, although potentially consisting of purely blind action, can develop to become highly regular and systematic in a mutually rewarding way.

Not all cases of rule-following involve only blind action. As mentioned before, following a rule in a material sense can require a certain number of steps of interpretation. Alice and Bob, in the example above, interpret the strategies that they wrote down on paper. But they don't, one presumes, need to further interpret the symbols used to write those strategies down. A step of blind action supports a further step of interpretation. Such complex processes can be explored using the signaling games framework. Barrett (2013b) gives an example of this. He proposes a model to capture the transitive behavior of pinyon and scrub jays. These birds can infer a linear order of colored elements by being trained on pairwise similarities between those elements. Barrett suggests that this could be explained by the evolution of two systems: a basic system that can learn to classify two stimuli in terms of their relative order, and a higher-order system that can in effect perform the transitive closure using the results of the basic system. Similar hypotheses about hierarchical processes could perhaps be explored further using models of multi-level selection or hypergames (briefly discussed in Section 4.3). These ideas are highly speculative, and their connection with Wittgenstein's rule-following considerations would need to be made much more explicit, but I see them as providing a glimpse to future research on the topic.

There is another important aspect of rule-following where the parallel with Lewis conventions can be informative. The discussion leading up to Remark 6.1.3 (p. 151) summarizes Wittgenstein's considerations about the normativity of rules. These try to address a common intuition: that there is a sense in which rules "ought to" be followed. This can ramify into a number of other ideas, fueled by the ways we talk about rule-following (Remark 6.1.4, p. 153). Wittgenstein discusses one way in which normativity takes shape: the idea that, given an expression of a rule, there is a correct way of following it. In particular, he dispels the thought that the criteria of correctness for the application of a rule could somehow be contained in it.

Not many positive remarks are given on normativity, but we can try to surmise a sketch of a more positive account from Wittgenstein's other remarks. They, I believe, can give us hints on how to conceive of the normative aspect of rule-following without falling prey to the usual problems. First, as mentioned before, our feelings of correct and incorrect applications of rules come from our familiarity with them. Via training (Remark 6.2.2, p. 156), we internalize certain ways to react to, or interpret, rules expressed as signs or orders. We end up following them blindly and thus often find it difficult to even conceive of alternative reactions or interpretations of familiar rules. Making use of a chart by associating the elements horizontally, following a signpost in a certain direction, identifying whether something is or is not a cube, continuing a number series based on the formula, interpreting a pointing gesture, pronouncing a letter or word, and so forth, all of these activities strike us as already containing their own criteria of correctness because of how one particular way of performing them was inculcated upon us. As following a rule is a custom (Remark 6.2.3, p. 158), these feelings are further reinforced by us observing others participating in those practices, or by participating in them ourselves. The same characteristic expressions of agreement, rejection, expectation, or encouragement used by a tutor to train a pupil in following a rule, are continuously used throughout our rule-following practices. Not only can we observe the regularities in the rule-following behavior itself, but also in its correction or reinforcement (what one could call the meta-behavior). A way to further try to clarify these ideas is to look at how similar issues arise in the context of Lewis' notion of convention.

There are two ways in which Lewis sees conventions as a "species of norms" (1969, pp. 97-100). The first has to do with instrumental rationality. Behaving in accordance to a convention is, given the structure of the coordination problem and the expected conformance of others, the choice that optimizes an agent's expected utility. Given that the latter represents, in Lewis's interpretation, the agent's preferences, and assuming that the agent is instrumentally rational, one could say that he "ought to" conform since using the strategy established by the convention is the choice that best answers directly to his preferences. Consider Alice and Bob again. Having agreed explicitly on a code to coordinate their next meeting place, and having no reasons to believe Bob would not conform to the signaling convention, it seems that Alice ought to follow the agreed-upon strategy since it is supposedly in her own interest to meet Bob.

The second way in which conventions have an element of normativity, according to Lewis, has to do with the expectation of social retaliation. Given that conformity to the convention answers not only to the agent's own preferences, but also to the preferences of others involved, it seems reasonable to expect that failure to conform is likely to be met with some form of punishment. If Alice flouts the convention and sends Bob the signal she was supposed to send when interested in going to Café One, but actually goes to Bistro Two, Bob might feel offended to have been stood up and refuse to meet Alice again. Alice ought to abide by the convention if she wants to avoid this kind of potential consequence.

The problem with these two ways in which some normativity can be argued to be present in Lewis conventions is that they arise from sources external to the framework (Guala, 2013). The agents' instrumental rationality is simply assumed by most game theory models. Although, as was mentioned in Section 2.2, some models that relax that assumption have been proposed, agents are always seen as striving for utility maximization even when their limited cognitive resources or other external factors prevent them from fully reaching that goal. Because it is built into the models by design, it is not something that can be studied by means of them but needs to be motivated for independently.

Regarding the second way in which Lewis sees normativity in convention, the mechanism of retaliation, although it is intuitive and recognizable in our everyday practices, it is presented in a strictly hypothetical sense and not actually incorporated into the framework. There is, therefore, no way to study its impact on enabling or maintaining conventions. However, there is no reason why this mechanism couldn't be captured. In fact, there is work in game theory that explicitly models it in the context of speculation about the evolution of cooperation (*e.g.* Axelrod, 1986; Andrews, Thommes, and Cojocaru, 2015). Although this literature is related to competitive problems captured in terms of the Prisoner's dilemma, one could adapt the idea to apply it to Lewis-style coordination problems. I am not familiar with any work in the literature that does so, but it seems like an interesting topic for future work. Whether or not it gives any insights into normativity in natural language remains thus to be seen.

I have been so far discussing the issues of rule-following in general. Wittgenstein's main interest in the topic is, however, not a general theory of what following a rule is. It is rather the implications of those considerations for questions relating to language and meaning. Rules play an important role in how we conceive of language learning and use. Especially since the advent and proliferation of formal education systems, we have a picture of learning a great deal of language by being taught explicit rules. We learn rules on how to pronounce and identify certain sounds, rules on how to combine them into words, and those into sentences. We learn further rules on how to convert all of those into written text, rules of spelling, hyphenation, capitalization, punctuation, and so on. This creates an everyday picture of language as governed by rules. The mainstream scientific view of these aspects of language is also similar in spirit, even though naturally much more technical[8]. For example, Hauser, Chomsky, and Fitch suggest that "we can profitably think about language as a system of rules placed within a hierarchy of increasing complexity." (2002, p. 1577) Although Wittgenstein's considerations on rule-following could probably help shine a different light on this picture (*e.g.* Waller, 1977), I will not be concerned with rules relating to phonology, morphology, or syntax here.

Wittgenstein's main underlying target has rather to do with particular intuitions regarding meaning. One way to see meaning as governed by rules is to see it in terms of the correspondence picture. If linguistic expressions have meanings, one would expect there to be rules on which linguistic expressions correspond to which meanings, so that one could associate the two in a systematic way. Both language production and understanding could then be imagined as processes that are guided by these rules. This picture goes along with a strong perspective on normativity. There is a right way of using or understanding a linguistic expression, by pairing the expression with its correct meaning (the meaning determined by the rule), and a wrong way, by failing to do so. Most problems with this picture were discussed in Chapter 5, but there are additional ones that are best seen in light of Wittgenstein's remarks on rule-following.

One important issue has to do with the determinacy of meaning. Talk about linguistic expressions having meanings goes together with talk about what *the* meaning of a word or a sentence is. Many theories of meanings qualify these statements in various ways (see Speaks, 2018). One could say that the linguistic expression by itself is not sufficient to determine its meaning, it needs to be paired with a context of utterance. Perhaps this does not yet determine meaning, but only an intension which needs to be evaluated in each particular circumstance. Perhaps it is not only

[8] In order to see this, it suffices to browse through any introductory text on modern linguistics, *e.g.* Akmajian et al. (2017).

an intension, but a Fregean content which determines a Russelian content which in turn determines an intension (Speaks, 2018, Figure 4). Whatever the number of additional elements and levels of indirection added, all of these theories share the idea that, given enough of those, one can determine the meaning of a linguistic expression.

In this respect, I side with Kripke's interpretation of Wittgenstein as presenting a case against the idea that there is ever an objective fact of the matter as to the meaning of a linguistic expression (Kripke, 1982). As summarized in Remark 6.1.2 (p. 149), rules are always up for interpretation. This includes rules associating an expression directly with its meaning, or with its character, or associating a character and a context to a content, or an intension and a circumstance to a referent, or what have you. Even though when we follow a rule we do so blindly (Remark 6.2.2, p. 156), that does not mean we could not have potentially acted otherwise.

With this in mind, I would like to return to some attempts at defining meaning in the context of the signaling games framework beyond those discussed in Section 5.4. Lewis (1969, pp. 143-152) proposes that we can extend the idea that a signal in a signaling system can conventionally mean either that a state holds, or that a certain action should be performed, to think of signals as indicative or imperative. Lewis characterizes this in terms of how much discretion the strategies allow the respective agents. If the sender strategy does not give the sender freedom to deliberate about which signal to send given the state that holds, but the receiver strategy gives the receiver freedom to deliberate about which action to perform when receiving a certain message, the signal is *indicative*. If the sender strategy gives the sender freedom to deliberate, but the receiver strategy does not give the receiver freedom to do so, the signal is *imperative*. If neither strategy is discretionary, or if both are, the signal is said to be *neutral*.

Consider again Lewis' example of the coordination problem between Paul Revere and the sexton of the Old North Church (Lewis, 1969, pp. 122-125). A non-discretionary receiver strategy could specify, for example, that "If one lantern is observed hanging in the belfry, warn the countryside that the redcoats are coming by land". A discretionary variant could say something along the lines of "If one lantern is observed hanging in the belfry, do whatever seems best on the assumption that the redcoats were observed setting out by land" (1969, p. 145). The idea is that the latter gives Paul Revere the freedom to determine his own actions, whereas the former does not. The signal of hanging one lantern in the belfry would have an imperative character in the first case, but not in the second. The main problem with this suggestion is that, if we see strategies as rules in a material sense, Remark 6.1.2

(p. 149) must lead us to see *all* strategies as ultimately discretionary. Paul Revere can take the first instruction and, as with the second, do as he pleases when observing one lantern in the belfry. Even though it would feel strange if he would warn the countryside that the redcoats are coming by sea, the rule can in principle always be interpreted otherwise than what seems familiar to us.

Could a descriptivist account make a difference? This is, to a certain extent, what Lewis actually has in mind. The distinction was originally made in terms of how one can "properly describe" a signal as a signal-that, meaning that a state holds, or as a signal-to, meaning that a certain action should be performed (1969, p. 144). The description is thus a third person perspective on the system. Huttegger (2007b) extends these notions of indicative and imperative meaning to agents with lower cognitive capacities by incorporating deliberation, as "any mechanism that processes information inputs and eventually leads to an output" (2007, p. 410), into an evolutionary signaling game. Zollman (2011) revisits this and proposes yet another way to flesh out this distinction using a model with multiple receivers. In my opinion, none of these attempts succeed. By taking a descriptivist stance towards strategies, they leave the task of finding a description of the meaning of signals up to the third party observer. What they need, in order to make the distinction between indicative and imperative meaning, is for there to be situations where a signal can *only* be described in terms of either indicating that a state holds or commanding an action to be performed. The problem is that neither the strategies, nor the larger context of the game can strictly determine the signals interpretation. If we see strategies as rules in a descriptivist sense, Wittgenstein's maxim (Remark 6.1.2, p. 149) again directly applies, but now to the third party observer who is to interpret them in the context of the game: where one sees an indicative signal that lets the receiver deliberate, another can see it as an imperative signal to deliberate.

Zollman is well aware of the problem[9] when it comes to the proposals of Lewis and Huttegger (Zollman, 2011, pp. 161-162, 164). He advances his own model to try to address the issue, but ultimately needs to acknowledge that perhaps the problem is not solved after all:

> It seems clear that in this case the plausibility of the two translations is stretched further than it was before, but not that those translations are somehow impossible. However, those who would point out that we have not eliminated the possibility of devising both indicative (or assertive) and imperative (or directive) translations of Sig A might have difficulty

[9] Zollman (2011) talks about what I have been calling interpretations of the meaning of signals as translations into English.

> describing what would be necessary to demonstrate that such a real distinction exists. (2011, p. 168)

The burden of proof is shifted to those who would like to demonstrate a real distinction between indicative and imperative meaning, but are bothered by the possibility that different interpretations seem to always be possible. Such a possibility undermines the distinction because it purports to be a "real distinction", *i.e.* objective and independent of interpretation. But perhaps the conclusion should rather be that the apparently unshakable problem that a signal (or a whole strategy, or a whole signaling game model) can always be interpreted otherwise just shows that the distinction these authors are seeking cannot be made *a priori* in an objective way.

The difficulty in upholding the distinction has to do with the lack of determinacy of meaning, and this is a direct consequence of the problem of interpretation raised by Wittgenstein. There is no rule that forces an interpretation upon us, and there is no linguistic expression that determines its own meaning. One might feel that a particular description of the use of a signal is more proper than others, as Lewis (1969, p. 144) puts it, or that one translation is much less plausible than another, as Zollman (2011, p. 168) puts it, but these judgments are vague (no characterization of what makes a description proper, or a translation plausible, is given), subjective, and based on intuitions and feelings of familiarity. The distinction is something that an individual can make, but it is not in any way an objective property that a signal, a signaling system, or a language has independently of interpretation.

In the previous sections I mentioned three conceptions of rules, but here I talked so far only about the material and the descriptivist. The reason for ignoring the third one is that the justificatory conception of rules does not have a good parallel in the signaling games framework. First of all, it would only make sense in interpretations of strategies as representing individual agents rather than populations. Second, models are typically (if not always) designed with a third person perspective in mind. There is therefore no room to capture what an agent would give as a justification for their behavior if asked for it. It would make sense that a rational agent using strategies as instruments would appeal to them in such a situation, but such a behavior is not spelled out in any model that I know of in the literature.

In conclusion, I argued that signaling games can help illuminate some of Wittgenstein's positive remarks on rule-following, in particular his notion of blind action and his appeal to custom and practice. The connection between the two can additionally inspire more detailed proposals on how to account for the connection between blind action and interpretation, as well as possibilities for modeling normativity in natural language. Wittgenstein's more skeptic remarks are also important for keeping us from falling prey to some intuitive urges with regards to interpretations of the signaling games framework. In particular, they should avert us from conceiving of meaning as determined by rules.

Conclusions

In this thesis, I tried to explore some connections between Wittgenstein's later philosophy and the framework of signaling games. I focused on the four major themes in the *Philosophical Investigations* that seemed more relevant to the task: Wittgenstein's remarks on method, his picture of language, his criticism of meaning as a form of correspondence, and his remarks on rules and rule-following. The objective was to shine light in two directions. First, if signaling games provide a way to study language that is largely in line with Wittgenstein's later philosophy, this insight can direct those who appreciate the latter to a more systematic set of tools that can help push it forward. Second, those who already appreciate the usefulness of signaling games for exploring hypothesis about language use can benefit from Wittgenstein's critical insights into one's philosophical presuppositions. This can further be relevant to the interpretation and use of signaling game models. I hope that I succeeded in showing that there is indeed a close fit between Witgenstein's later philosophy and the framework of signaling games, and that there are lessons to be drawn in both directions regarding each of the four major themes addressed.

In Chapter 3, I delved into issues of methodology in philosophy. I argued that Wittgenstein's negative remarks are directed not against particular methods, but rather at an idealizing[10] attitude that can send philosophers in pursuit of chimeras. I also defended that an alternative pragmatist attitude can be surmised from Wittgenstein's remarks on explanations of meaning, and his own use of toy language-games. One can further see many similarities between the latter method and the use of models in the signaling games literature. However, the risk of taking an idealizing attitude always looms. A method (especially the history of its use) can promote a different attitude, but it cannot enforce it. That is why it is important to understand Wittgenstein's warnings and always keep a pragmatist attitude in mind when making use and interpreting signaling game models. The signaling games framework provides useful tools to better understand the complexities of communication, but it should not be taken as a theory of meaning, or be used in a way that falls prey to the urges of the idealizing attitude.

Chapter 4 characterizes the picture of language and meaning that I argue is endorsed in the *Philosophical Investigations.* Wittgenstein describes language in terms of practice, drawing attention to the involvement of agents, the interconnection with other activities, the importance of purpose, and its heterogeneity and dynamism.

[10] In an everyday sense, not to be taken to refer to the schools of thought in philosophy commonly known as idealism.

In the context of this picture, I defend that it is best to interpret Wittgenstein's remarks on use as methodological advice: if you want to better understand meaning, investigate how linguistic expressions are used. I argued that the signaling games framework allows one to embrace this picture of language quite well, by incorporating agents, actions, and purpose by design. Most approaches within the framework promote taking heterogeneity and dynamism into account, some models even help illuminating those aspects of language further. Considering Wittgenstein's remarks additionally reminds us of the open-ended nature of research into language and meaning. Signaling games are still, in many ways, very limited tools in the face of the complexity of the phenomena, but the flexibility of the framework allows one to embrace this and see interesting avenues for future work.

Chapter 5 deals with one of Wittgenstein's major grievances against most prevailing conceptions of meaning in philosophy. The focus is on rejecting the intuition, rooted in everyday talk, that linguistic expressions have meanings, like entities that they carry around or stand in correspondence to. This intuition constitutes what I called a correspondence picture of meaning. I explored the arguments leveraged against three incarnations of this idea: externalism, logical atomism, and internalism. I then discussed some ways in which this intuition has slipped into the signaling games literature, in particular in attempts to define signal meaning in those models. I argued that these proposals are problematic, especially in light of Wittgenstein's remarks. When using the signaling games framework one can, and should, focus on signal use and "leave meaning to look after itself" (Lewis, 1969, p. 122).

Another major topic in the *Philosophical Investigations*, which I address in Chapter 6, is rule-following. Wittgenstein's remarks on the topic touch upon various misconceptions surrounding the role of rules in shaping our behaviors that exhibit regularities. Criticism is raised against the ideas that rules are sufficient or even necessary to explain all of those behaviors, that they can be a source of normativity, and that they have the power to determine actions or outcomes by themselves. I discussed the implications these remarks have on interpretations of strategies in signaling games, to what extent there is normativity in those models, and whether or not one could say that meaning is determined. Some upshots of that discussion are suggestions for future research, for example in linguistic normativity. Others constitute further criticism of conceptions of signal meaning in deterministic, rule-based terms.

Overall, my main conclusion is that the signaling games framework fits Wittgenstein's later philosophy of language quite well. It is compatible with his metaphilosophy and methodological practice; it embraces a heterogeneous, dynamic, practice-based picture of language; and it can avoid the problems related to correspondence-based and rule-based conceptions of meaning. Some mismatches point to avenues for future research, like exploring multi-population dynamics and hypergames, or developing signaling game models of linguistic normativity. Some of Wittgenstein's remarks serve as reminders to be cautious about certain interpretations and uses, like taking the framework as constituting a theory of meaning, or trying to provide some definition of signal meaning.

Signaling games are of course limited, in the way all tools are. There are technical issues, theoretical blind spots, and a lot of room for future work. Signaling game models mostly focus on very basic linguistic interactions. As such, the approach can be taken as oversimplifying language. I believe this would be a misinterpretation. Despite the general focus on basic interactions, no one denies our ability to construct complex grammatical sentences or use language in highly strategic ways. And some of the research briefly discussed in Chapter 2 shows that the framework can accommodate for modeling some of these kinds of interactions. In keeping with the pragmatist attitude advocated in Chapter 3, one should not forget that each model is tailored to a particular language-game. It would be a mistake to overgeneralize a particular model as being representative of language as a whole. The framework cannot, and need not, capture everything there is to be said about language. There is only a problem if one believes otherwise. I personally think that it is important to first understand the most simple language-games well before modeling more complex ones, since the latter build on the former. Problematic intuitions about meaning can easily creep in if not kept in check. Wittgenstein's remarks can play the role of reminders against such urges.

A cornerstone of the *Philosophical Investigations* is the observation that when we talk about language we are usually talking about a multifarious and highly complex phenomenon, a product of the interaction of many moving parts. A contemporary picture of language that I think is very much in line with Wittgenstein's later philosophy is summarized in the following passage:

> The system consists of multiple agents (the speakers in the speech community) interacting with one another. The system is adaptive; that is, speakers' behavior is based on their past interactions, and current and past interactions together feed forward into future behavior. A speaker's behavior is the consequence of competing factors ranging from perceptual constraints to social motivations. The structures of language emerge from interrelated patterns of experience, social interaction, and cognitive mechanisms. (Beckner et al., 2009)

This is what the authors call seeing language as a *complex adaptive system*. To embrace this picture is to reject the temptation to reduce language to some static well-delimited object that can be investigated in isolation. In order to study the emergent patterns generated by such a complex adaptive system, one needs tools that take all those elements into account and fully embrace both spatial and temporal aspects of their interrelations. For various reasons argued throughout this thesis, I believe that the signaling games framework provides such tools and promotes a study of meaning along those lines.

Additionally, in order to gain more knowledge of such a complex phenomenon as language, it is important to embrace a multidisciplinary approach. The signaling games framework draws inspiration and makes use of concepts from areas such as economics, biology, psychology, mathematics, physics, and others, depending on the issues at hand. The philosopher that wishes to better understand language and meaning needs to be able to reach beyond their field and explore notions from other areas. This goes hand in hand with a pluralistic methodological stance. The literature summarized in Chapter 2 makes use of classic game-theoretical notions, like for example the Nash equilibrium, tools from evolutionary biology, like equations of population dynamics, general mathematical tools, like graphs representing different network structures, knowledge from psychology and artificial intelligence, like models of learning mechanisms, notions from physics, like deterministic chaos, measures from information theory, like entropy and other metrics to help characterize systems at a higher level of abstraction, and many others. One should also not refrain from integrating insights from anthropology, applied linguistics, sociology, experimental psychology, and other related areas. All of the areas of knowledge mentioned and their tools can hypothetically be relevant for the task of the philosopher, and none should be off limits. Philosophy is inevitably an armchair enterprise, but it should not be a nearsighted one.

Part of Wittgenstein's aim in the *Philosophical Investigations* was to open the eyes and minds of philosophers that, like his early self, tend to underestimate the complexity of the phenomenon of language. This involved undermining some reductive assumptions and intuitions that formed the foundations of the systematic approaches to meaning in his purview. But one need not shy away from systematicity in order to reflect on language and meaning along the lines of Wittgenstein's later philosophy. I hope to have shown that the framework of signaling games serves as a good example that one is not incompatible with the other. The similarities between the two pictures of language, and the added power of mathematical formalization and computer simulation that the signaling games framework brings to the table, make the latter an attractive approach for studying language and meaning that is both systematic and in line with the *Philosophical Investigations.*

Appendix A

Towards an ecology of vagueness[*]

Abstract

A vexing puzzle about vagueness, rationality, and evolution runs, in crude abbreviation, as follows: vague language use is demonstrably suboptimal if the goal is efficient, precise and cooperative information transmission; hence rational deliberation or evolutionary selection should, under this assumed goal, eradicate vagueness from language use. Since vagueness is pervasive and entrenched in all human languages, something has to give. In this paper, we focus on this problem in the context of signaling games. We provide an overview of a number of proposed ways in which vagueness may come into the picture in formal models of rational and evolutionary signaling. Most argue that vague signal use is simply the best we can get, given certain factors. Despite the plausibility of the proposals, we argue that a deeper understanding of the benefits of vagueness needs a more ecological perspective, namely one that goes beyond the local optimization of signaling strategies in a homogeneous population. As an example of one possible way to expand upon our current models, we propose two variants of a novel multi-population dynamic of imprecise imitation where, under certain conditions, populations with vague language use dominate over populations with precise language use.

[*] The contents of this appendix are a reproduction of the publication with the same title by Correia and Franke (2019), with potential very minor adjustments. Part of Section A.2 served as the basis for Section 2.1 and contains some repeated content, but given that it is also interspersed with remarks not included there, I reproduce it here in its entirety.

A.1 Vagueness and rationality

The classical philosophical problem of vagueness is most starkly embodied by the sorites paradox. The original formulation is attributed to Eubulides, an ancient Megarian philosopher (Sorensen, 2009), and uses the example of a heap of sand: if no removal of one grain of sand can make a heap into a non-heap, one can repeatedly remove all but one grain of sand from something that is clearly a heap and be forced to acknowledge that the remaining single grain of sand is still a heap; otherwise, it seems, one would have to accept that there is a determinate number of grains that forms a heap, and anything under it is not a heap. Neither choice is, however, intuitively satisfying. The paradox is interesting because it can be made general and re-applied to many other words besides 'heap'. Predicates for which one can find a suitable instance of the general formulation of the sorites paradox are called *vague*. Paradigmatic examples besides 'heap' include 'tall', 'red', 'bald', 'tadpole', and 'child' (Keefe and P. Smith, 1999). How widespread is the problem? It is easy to find more examples of predicates based on more finely grained properties—as 'tall' is intuitively based on height—for which constructing a sorites paradox would be easy. Mereological nihilists argue that instances of the sorites paradox can be designed for any material object that can be decomposed into small enough parts. If one subscribes to the scientific picture of matter as composed of molecules and atoms, this applies to tables and chairs, cats and mats, and any other ordinary thing (Unger, 1979). Bertrand Russell famously argued (1923) that all words, including "the words of pure logic," are vague when used by human beings.

If one thinks of language as governed by logical rules, and of rationality as including the ability to follow those rules, the sorites paradox seems *prima facie* to demonstrate that vagueness and rationality are incompatible. But one can think of language in different terms. One possibility is to think of signs as tools agents use to coordinate actions. An example of a way to formally study language along those lines is the game-theoretic framework of signaling games (Lewis, 1969). Within such an approach, rationality can be seen as the ability to choose the use of signs that is optimal to achieve some form of coordination. The existence of vagueness in such models would not be at odds with rationality as long as vague languages turn out to be optimal for the purposes at hand. However, as Barton Lipman (2009) argues in detail, this is typically not the case. The problem can be put very succinctly as follows. In standard game-theoretic models of communication, vague signal use yields a lower expected utility than crisp use. Therefore, given that the dynamics (be it natural selection, cultural evolution, or rational choice) maximize utility, vagueness should be weeded out by these forces, giving rise to only precise languages. Thus it would be irrational to stick to vague language use when one could (theoretically) switch to a better system. But vagueness is pervasive in natural language, and there is no reason to believe it is going away. The problem seems to be a theoretically serious one, but it does not obviously undermine our everyday linguistic practices. Most of the time we seem to communicate just fine. Therefore, the issue must lie with the conceptions we have of the forces or mechanisms that underlie those practices. Lipman concludes that "we cannot explain the prevalence of vague terms in natural language without a model of bounded rationality which is significantly different from anything in the existing literature" (2009, p. 1).

Here we survey proposals for addressing vagueness in the context of signaling games, but before we start we want to establish some vocabulary to better frame the upcoming discussion. Rationality is an elusive notion that is frequently debated in areas like philosophy, psychology, and economics. Discussions around it usually touch on different aspects of the concept without always clearly demarcating them; this is what we want to do before going further. The aforementioned logical picture of language and meaning is focused on dichotomies like true and false, meaningful and meaningless, correct and incorrect. A sentence can be true or false only if it is meaningful, and it is meaningful if it is constructed according to correct rules. Rationality is intimately connected with the ability to follow certain procedures, not only of sentence production, but ultimately of sentence combination and reasoning (think of what is required for making a logically valid deductive inference). This is an example of what we will call a *procedural* account of rationality, one which focuses more on the means rather than the ends. One can also do the opposite and focus on the consequences instead. *Instrumental* rationality, as it is typically called, characterizes an agent's choice as rational if it maximizes the possibility of achieving a desired goal, regardless of the means. The notion is linked to David Hume (1738), epitomized in the following assertion: "Reason is, and ought to only be the slave of the passions." Instrumental rationality is close to a notion of rational choice that is used in economics and game theory:[2] agents are rational if and only if they make decisions that maximize their expected utility. A more in-depth discussion of the opposition in the context of theoretical economics can be found, for example, in the work of Herbert A. Simon (1986).[3]

When developing models where rationality is relevant, be it constructing a logical system or setting up a formal game, the assumed epistemic relation between agents

[2] In fact, it has been argued (Vanderschraaf, 1998) that Hume's whole account of convention is very closely in line with modern game theory.

[3] Simon uses the term substantive instead of instrumental rationality, but the characterization is basically the same (Simon, 1986, pp. 210-212).

and their environment can come to bear on considerations of rationality. The verdict over how rational certain choices are can vary depending on how accurate and complete an agent's knowledge is of its environment, the goals to be achieved, the choices or rules available, the relation between those choices and the objectives, and so forth. In this respect, we will call an agent *omniscient* if it is in possession of the same information as the modeler, whereas of an agent with less than that we will say that it only has *limited* awareness of the relevant aspects of the model. Models working within the logical picture typically do not make a distinction between modeler and agent, and thus lack room to express these epistemic gaps. By abstracting away from language users, these models also typically do not represent potential interactions between them, let alone allow for repeated interaction and language change. In other types of models, however, a further aspect of rational choice needs to be considered, namely the ability of agents to make accurate predictions about how other agents behave. We can say that an agent is more or less *strategic* depending on the extent to which she is able to anticipate the actions and beliefs of other agents, and to predict medium/long term gains from repeated play. Lack of perfect knowledge of the situation or lack of ability to choose strategically can be caused by many possible factors. These include, among other things, limitations in handling information (receiving, storing, retrieving, transmitting) and limited computational resources to solve complex problems. We can talk about *bounded* rationality to characterize the choices of such agents. While the above definition of instrumental rationality is usually understood as requiring a single choice to maximize the expected utility in a single concrete decision situation, one might also be interested in more general choice mechanisms (Zollman and Smead, 2010; Hagen et al., 2012; Fawcett, Hamblin, and Giraldeau, 2013; Galeazzi and Franke, 2017). A choice mechanism is a general way of behaving for an agent involved in a variety of decision situations. When considering only a single situation, rationality can only be *local*. If we take into account the possibility of the agent's choice or choice mechanism hinging on multiple situations, we can also talk about *global* rationality. This can be important because, hypothetically, there could be choice mechanisms that are sub-optimal at a local level (for each situation) but are actually perfectly rational at a global level.

Note that we consider that, in theory, most combinations of these aspects are possible. Although we have been pinning procedural rationality to the logical picture of language, this is only with the most traditional logical systems in mind. We are not denying that advances in dynamic, epistemic, fuzzy, paraconsistent, and other types of logic could potentially enable one to capture procedural rationality with different characteristics. Game-theoretical models of language can, on the other hand, also combine local instrumental rationality with omniscient highly strategic agents. Our objective here is not to survey all the possibilities. We want to focus on signaling games as a framework for the study of language use and meaning. The vocabulary just introduced will, we hope, help inform the discussion that follows. We proceed by introducing the framework of signaling games in section A.2. In section A.3 we look into explanations of vagueness in a particular kind of signaling game. Section A.4 tries to generalize the considerations of these proposals to argue for an approach to vagueness anchored in a more global notion of rationality. We propose and analyse the results of a novel multi-population model of imprecise imitation in section A.5, and summarize our conclusions in section A.6.

A.2 Signaling games

Signaling games were first introduced as models of communication by David Lewis (1969). In order to support the idea that linguistic conventions can arise without any prior conventional activity, Lewis considers situations where agents' choices involve sending and receiving signals or messages.[4] One can think of two players with different roles. The first player, the sender, has knowledge about which of a number of possible states of affairs obtains and, depending on this information, chooses a signal to send. The second player, the receiver, has knowledge about which signal the sender chose and, based on this information, chooses one of several possible responses. A preference relation exists between responses and states of affairs, and a payoff is attributed to each player based on the choices of both. Note that Lewis assumes that no player has any preference regarding the particular signal that is used for a given state, provided that it enables advantageous coordination with responses. Formally, in order to describe the setup all we need is to specify a set of possible states of affairs T, a probability measure P such that $P(t)$ is the probability or frequency with which $t \in T$ occurs, a set of available signals or messages M, a set of responses or actions A, and a pair of utility functions $U_{S,R} : T \times A \to \mathbb{R}$, one for the sender and one for the receiver, each of which yields a payoff value for each possible pairing of state and action. These so-called signaling problems can be seen as particular cases of coordination problems if we consider the players' choices to be of contingency plans or strategies. A sender strategy is a specification of a choice of message for each possible state of affairs. It thus describes the sender's behavior conditional on the state of affairs that obtains. A receiver strategy analogously

[4] These terms will be used interchangeably.

specifies a choice of action for each possible message. Formally, what the sender chooses is a function $\sigma : T \to M$ and the receiver a function $\rho : M \to A$. The expected utility EU of a strategy can be calculated by using the utility function and aggregating payoffs for all pairings of states of affairs and actions, weighted by the probability of each state. Concretely, the expected utility of σ given ρ is $\mathrm{EU}_S(\sigma \mid \rho) = \sum_{t \in T} P(t)\ U_S(t, \rho(\sigma(t)))$, and the expected utility of ρ given σ is $\mathrm{EU}_R(\rho \mid \sigma) = \sum_{t \in T} P(t)\ U_R(t, \rho(\sigma(t)))$. As an example, consider a game with $T = \{t_1, t_2\}$, $M = \{m_1, m_2\}$, $A = \{a_1, a_2\}$, $P(t_1) = P(t_2) = 0.5$ and the following utility matrix:

	a_1	a_2
t_1	$1,1$	$0,0$
t_2	$0,0$	$1,1$

Consider sender strategy $\sigma = \{t_1 \mapsto m_2, t_2 \mapsto m_1\}$ and receiver strategy $\rho = \{m_1 \mapsto a_2, m_2 \mapsto a_1\}$. These would have an expected utility of 1 for both sender and receiver, since when t_1 obtains with probability 0.5 the sender will use m_2, and to this message the receiver will respond with a_1, which achieves a payoff of 1, and when t_2 obtains with probability 0.5 the sender will use m_1, and to this message the receiver will respond with a_2, which also achieves a payoff of 1. They also represent one of the two stable conventions in this game, the other being the pair of strategies $\sigma = \{t_1 \mapsto m_1, t_2 \mapsto m_2\}$ and $\rho = \{m_1 \mapsto a_1, m_2 \mapsto a_2\}$. Conventions of this kind in a signaling problem are what Lewis calls *signaling systems*. An example of complete miscoordination would be $\sigma = \{t_1 \mapsto m_1, t_2 \mapsto m_2\}$ and $\rho = \{m_1 \mapsto a_2, m_2 \mapsto a_1\}$. Partial coordination is achieved, for example, by $\sigma = \{t_1 \mapsto m_1, t_2 \mapsto m_1\}$ and $\rho = \{m_1 \mapsto a_1, m_2 \mapsto a_2\}$.

Lewis' account of the stability of conventions rests on what could be considered strong demands. Namely, there needs to be a state of affairs that indicates to everyone involved that a certain regularity will hold, as well as "mutual ascription of some common inductive standards and background information, strategic rationality, mutual ascription of strategic rationality, and so on" (1969, pp. 56–57). Agents are thus envisioned as omniscient and highly strategic. These requirements can seem excessive, and even more so if we consider how simple signaling systems are when compared to human languages. The models were introduced to help explain how language could have gotten off the ground as a conventional system without any sort of prior agreement. However, if one considers the circumstances of the origins of human language, it seems implausible that the agents that started making use of primordial signaling systems which (hypothetically) evolved into languages possessed such advanced rationality. Furthermore, communication through simple message exchange is something that almost all animals do: monkeys use calls, birds use singing, bees use dances, ants use pheromone trails, and so on. A plausible account of the origin of language should first explain how signaling systems could get started, without requiring high standards of rationality from the agents involved.

In order to address this problem, Brian Skyrms (1996) proposes studying signaling problems in evolutionary terms. Rather than imagining, as Lewis does, rational agents making conscious decisions in possession of knowledge of the game and expectations of the behavior of other agents, one can imagine a simpler scenario inspired by biological evolution: there is a population of agents with biologically hardwired behaviors for engaging in interactions characteristic of a signaling problem; utility does not represent preference, but rather fitness for survival and reproduction; the make-up of the population evolves based on the relative fitness of the strategies represented in the population. Such a setup attempts to capture the main features of natural selection: in a diverse population, agents with more successful strategies thrive, while agents with less fit strategies die off. Although the inspiration for this scenario is biological evolution, similar things could be said about how ideas spread in a population of agents who can adopt or abandon them depending on how successful they prove to be (Benz, Jäger, and van Rooij, 2006; Pagel, 2009; Thompson, Kirby, and K. Smith, 2016), *i.e.* we can interpret these notions in terms of cultural evolution (Dawkins, 1976; Boyd and Richerson, 1985). The principles can be captured in formal models that abstract away from details of single interactions and behavior of individual agents, for example in the replicator dynamics (Taylor and Jonker, 1978). The only things relevant to this equation are the relative proportions of strategies in a given population and the utility function. Using it, one can compute which strategies evolve under which conditions.

Skyrms' evolutionary game theory approach to signaling games not only gives more plausible grounds to support Lewis' discussion of convention, it also accomplishes an important conceptual change: it moves most of the theory and mathematical formalism to the descriptive side of the investigation. Utility represents how the modeler views the signaling problem and understands the relative advantages or disadvantages of different possible strategy combinations. Dynamics describe how strategies can evolve when driven by mechanisms of utility maximization. The shift in perspective allows interpretations that accommodate limited non-strategic agents. While the general framework manages to abstract quite some details away from the formalization, it nevertheless leaves room for them, especially when it comes to the dynamics. We have already mentioned the replicator equation that can be seen as representing biological or cultural evolution, but one can also use dynamics inspired by learning mechanisms (Roth and Erev, 1995), or even ones assuming a high degree of knowledge of the game and other players (Gilboa and Matsui, 1991; Mühlenbernd, 2011; Spike et al., 2017). This range of options goes hand in hand with a range of pictures of rationality, from nothing more than survival of the fittest in a biologically-inspired setting, to a certain degree of instrumental but limited and non-strategic rationality in the case of learning dynamics, to higher levels of rationality and even recursive strategic reasoning about the co-players' beliefs and choices. Each of these can be utilized depending on the problem that one is interested in characterizing. Thus, although Skyrms shows that high requirements of rationality are not necessary for signaling conventions to evolve, the framework does leave room for the study of linguistic interactions between highly strategic agents.

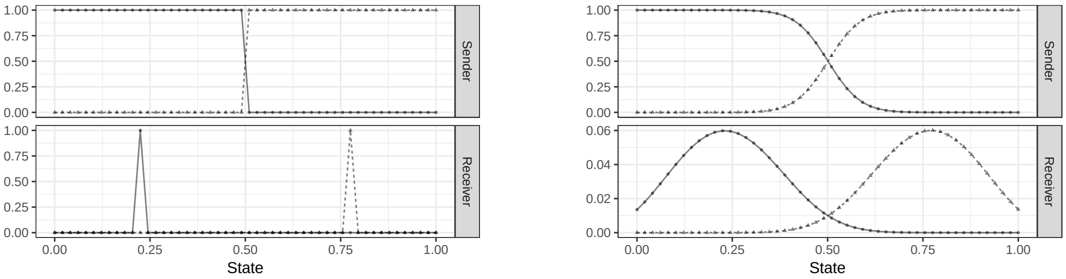
The characterization of signaling problems in terms of evolutionary game theory allows us to explain why certain equilibria come to be and how. A core notion in this context is that of an *evolutionary stable state* (Maynard Smith, 1982): an equilibrium situation that a population tends to under standard evolutionary pressures, and to which it returns if slightly disturbed. With these tools, one can better understand why signaling systems are stable even without any strong assumptions of rationality. One can also map out which initial conditions drive the system towards which equilibria and which do not. In a simple case like the example discussed above, an evolutionary process of the kind described always drives the population into a state where one signaling system takes over completely. More complex signaling problems may have different evolutionary outcomes, sometimes unexpected ones. Skyrms (2010) gives an overview of different topics studied using signaling games, including expansions of the framework itself (for example, considering other dynamics beyond the replicator equation), exploration of other factors that impact the evolution of signaling (for example, how agents are interconnected), or variations on the signaling problem and its basic assumptions (for example, loosening the alignment of interests in order to provide accounts of deceptive signal use). Other uses of signaling games include discussions of categorization (Jäger and van Rooij, 2007), compositionality (Barrett, 2009), incommensurability (Barrett, 2010), to name a few. More recent overviews are given by Huttegger (2014), Huttegger, Skyrms, Tarrès, et al. (2014), and Franke and Wagner (2014). In the following section, we discuss how a particular type of signaling game has been used to address the problem of vagueness.

A.3 Vagueness in sim-max games

The sorites paradox requires us to assume a relation between the vague terms and a more precise underlying dimension (height for tallness, number of hairs for baldness, number of grains of sand for “heapness”, and so on). Not only does this property need to be much more fine-grained than the vague term, but it also needs to have some structure: there is at least an order between the elements in it (thinking of height in centimeters, $180 > 179 > \ldots > 120$), and usually even a degree of how far apart these elements are from each other. In terms of signaling games, one can model this using a state space constituted of values of the underlying dimension, and a message space constituted by the terms in question. Because of the difference in granularity, we will typically be interested in cases where the state space is much larger than the message space. We can model the structure of the state space by defining a distance or similarity function between every value, effectively making it a metric space. Another important ingredient of the sorites paradox is the acknowledgment of a certain degree of tolerance with respect to whether a certain term applies or not. This tolerance decreases with distance in state space: assuming a 180cm person is tall, one would easily tolerate the use of the term for a person measuring 179cm, less so for someone who is 170cm, and much less so for 160cm. This can be modeled using a utility function that is continuous rather than discrete and that monotonously decreases with distance, *i.e.* success is not a matter of black and white, right or wrong, but a matter of degree, of how close the receiver got to the optimal response to the sender’s perceived state.

The simplest type of game to study in this scenario is one where the state space and the action space are the same. We can imagine this as a game of guessing states of affairs: the sender has knowledge of a particular state, sends a message to the receiver, who in turn has to guess it; their payoff, as discussed above, is proportional to how close the guess got to the original state. These games, called similarity-maximization or sim-max games for short, were first introduced by Gerhard Jäger and Robert van Rooij (Jäger and van Rooij, 2007; Jäger, 2007) and further studied by Jäger, Metzger, and Riedel (2011). What these authors find about this setup is that the evolutionary stable states are what they call Voronoi languages. Roughly, these are situations where the sender uses messages in a way that can be seen as partitioning the state space into convex regions, and the receiver responds with the central element of those regions.

In an abstract setup, using 50 states uniformly distributed over the unit interval and two possible messages, such an optimal language looks like what we see in Figure A.1a: at a specific point, the probability that the sender uses one message



(a) Crisp

(b) Vague

Figure A.1: Example strategies for a state space with 50 states. Each line corresponds to a message. For the sender, it plots for each state the probability that the message is used. For the receiver, it plots for each state the probability that the response is that state, given the message.

decays sharply from 1 to 0, and increases sharply from 0 to 1 for the other message; the response of the receiver for each message is a degenerate distribution over the state space which assigns all probability mass to a single state. These strategies give mutually optimal behavior for a case where the prior probability is the same for each state and utility is a linear or quadratic function of the distance between the actual state and the receiver's guess. In general, at which point the sender switches the use of messages and which guesses of the receiver are optimal critically hinges on the priors over states and the utility function. Still, confirming Lipman's argument, there is no vagueness in such optimal languages.

The sender/receiver strategy pairs that we are looking for look more like Figure A.1b, where the sender's probability of choosing a certain message gradually changes, and the receiver strategy assigns a positive probability to more than one state for each message. What characterizes vagueness in these models is thus a smooth and monotonous transition between parts of the state space where a message is clearly used and where it is clearly not. This means that, for some states in the middle of the state space, there is uncertainty as to which message will be used by the sender, whereas for states in the extremes this is as good as certain. For the receiver part, there is uncertainty as to which state will be picked for each message.

The interpretation of this uncertainty will be different depending on the interpretation of the model. If we see it as an explanatory model of how two agents play the game, we can see it as randomization. If we interpret the model as descriptive, it simply represents expected behavior in a manner agnostic to the underlying mechanisms. A third option is to see probabilities as capturing relative numbers in a population of agents. For example, if the sender strategy assigns a probability of 0.4 to the event of message m being sent for state t, this would mean that 40%

of the population uses that message for that state. This latter option leaves open the possible interpretation that each agent commands a crisp language and vagueness is only a population-level effect. However, given the level of abstraction of the description so far, none of this is necessarily implied by the model. Our preferred stance is to see the model as descriptive, but we do not wish to focus on that debate here. The question we want to address is: what additional modifications to sim-max games are sufficient for optimal languages to be more like Figure A.1b, rather than like Figure A.1a?

Franke, Jäger, and van Rooij (2011) make two suggestions as to how vagueness in a signaling game can be boundedly rational, *i.e.* how vagueness could arise as a consequence of cost-saving limitations in the cognitive capacities of instrumentally rational agents. The first proposal is called limited memory fictitious play (LMF) and models agents playing a sim-max game where their ability to recall past interactions with others is limited to a certain number. For a given interaction, each agent uses her limited memory of the other agent's past behavior to estimate the other's strategy, and plays an instrumentally rational best response to that strategy. In order to study the evolution of strategies in repeated interaction, the authors model several individual agents in actual play. What they observe is the emergence of vague signaling at the level of the population, *i.e.* population averages of individual strategies exhibit the hallmarks of a vague language. This is because each agent recalls a different history of previous play and so holds slightly different beliefs about language use. However, each agent still commands a crisp language, which is an inadequate feature of the model if the intention is to capture how vagueness presents itself in human languages.

In order to overcome this limitation, Franke, Jäger, and van Rooij make another proposal using the notion of a quantal response equilibrium (QRE). The idea, inspired by experimental psychology, is to model the choice of best response as stochastic rather than deterministic.[5] A prominent explanation for such soft-max or quantal choice behavior is that agents make small mistakes in the calculation of their expected utilities (Train, 2009). They still choose the option with the highest expected utility, but each assessment of the expected utilities is noise perturbed. This, in turn, may actually be boundedly rational since the calculation of expected utilities relies on assessing stochastic uncertainty, which in turn may be costly to calculate precisely. Choice based on a few samples from a distribution can be optimal if taking more samples or other means of better approximating probabilistic beliefs is resource costly (Vul et al., 2014; Sanborn and Chater, 2016). The de-

[5] Probabilistic choice rules are also the source of vagueness in recent accounts by Lassiter and Goodman (2017) and Qing and Franke (2014).

gree to which agents tremble in the calculation of expected utilities and therefore deviate from the instrumentally rational behavior can be characterized by a parameter. Franke, Jäger, and van Rooij find that for low values of this parameter, only babbling equilibria are possible, where sender and receiver simply randomize, respectively, message and interpretation choice uniformly. Above a certain value of the parameter, other equilibria of the kind described in the beginning of this section arise, where agents communicate successfully, though not perfectly, using fuzzy strategies similar to those depicted in Figure A.1b. However, it is not clear whether soft-max choices capture the right stochastic trembles in decision making as they would arise under natural sources of uncertainty about the context (see Franke and Correia, 2018).

Cailin O'Connor (2014) proposes a way in which vagueness could be expected to evolve as a side-effect of a particular type of learning process. She studies sim-max games driven not by rational choice dynamics, but by generalized reinforcement learning (GRL), a variant of Herrnstein reinforcement learning (HRL) (Roth and Erev, 1995). In HRL, agents learn to play a signaling game by strengthening particular choices (of messages for the sender, of responses for the receiver) proportionally to how successful those choices prove to be in an interaction. O'Connor's proposal is to model generalization as the propagation of reinforcement to nearby states, where "nearby" is defined in terms of distance in state space. For example, if a sender was successful in using message m for state t, she will not only positively reinforce that choice of message for t, but also for states similar to t. This is done to a degree that is proportional to the similarity between t and other states. The dynamics gives rise to vague signaling of the kind we are looking for.

Although there is a close relationship between reinforcement learning and population-level dynamics (Börgers and Sarin, 1997; Beggs, 2005), O'Connor's GRL is, on the face of it, an account of learning between individual players. Also, we need further justification for linking generalization in reinforcement so closely to the underlying payoff function of the sim-max game. Why should agents evolve to generalize in exactly the right way? O'Connor suggests that, despite a language with vagueness having lower expected utility than a language without it, the learning mechanism that induces vagueness does have evolutionary advantages: it achieves higher payoffs in a shorter period of time. From a global point of view, learning speed can be an advantage. Imagine an initial population of agents with random strategies, some using GRL and others using classical HRL. Although the latter type of agent can hypothetically develop a precise and more efficient signaling system, agents using GRL could coordinate on vague signaling strategies with high (though not optimal) expected utility sooner than agents using HRL. In such a scenario, they could drive the other agents to extinction before the latter had time to achieve coordination and reap the benefits of a more precise signaling system.

A similar finding is made by Franke and Correia (2018) when studying a variant of the replicator dynamics in which individual agents do not have the ability to generalize but simply make perceptual mistakes. In a scenario where agents learn by imitation, if one assumes that they do not have perfect perception, there will always be the possibility for senders to confuse states and thus learn associations with messages that are different than the ones observed, and for receivers to similarly mix up responses. Furthermore, it seems reasonable to assume that this confusability is proportional to state similarity, *i.e.* that the more similar two states are, the more likely it is that they will be mistaken for each other. Incorporating these considerations into a derivation of the replicator dynamics based on imitation processes, Franke and Correia develop a variant of the dynamic that also induces vague signal use of the kind we expect here. The consequence is very similar to that of the GRL model discussed above, in that the way the behavior for a given state is updated takes into account the behavior of similar states, proportionally to their similarity. Given the known relation between reinforcement learning and the replicator dynamics (Beggs, 2005), it is actually quite plausible that the two are tightly related (although this would need to be formally demonstrated). The account is, furthermore, interpretable at a lower level of rationality.

The motivation underlying this model of vague signaling is still one of inevitability. A vague strategy is not claimed to have higher expected utility than a crisp one. However, the authors observe an effect similar to that pointed out by O'Connor: signaling converges faster and more often in scenarios where there is some degree of state confusability. Furthermore, they observe one additional potentially beneficial property. Running several rounds of simulation for each parameter set, they measure for each group of results how close resulting strategies are to each other, and how they would fare playing against one another. The results show that the within group distance between strategies becomes smaller with growing confusability, and the within group expected utility is actually higher for strategies evolving under a certain degree of state confusion. Thus, some amount of uncertainty seems to promote more homogeneous populations of signalers that are better at achieving cooperation within a group. What is left to show, as it was with O'Connor's GRL approach, is whether the potential payoff advantages that were observed in simulations actually suffice to promote vague language use in an encompassing model of multi-level selection. The following section motivates the need for taking a more ecological approach to the evolution of vagueness, before section A.5 gives a concrete model.

A.4 The ecology of vagueness

Arguments of the kind presented by Lipman (2009), that vague signal use is suboptimal when compared to crisp use, work under various assumptions. Part of the picture formed by these assumptions is a highly idealized conception of the agents involved and of the context in which they develop and use signals. These idealizations probably originate, via game theory, from the conception of rationality of traditional theoretical economics. Herbert A. Simon describes this picture as follows:

> Traditional economic theory postulates an "economic man," who, in the course of being "economic" is also "rational." This man is assumed to have knowledge of the relevant aspects of his environment which, if not absolutely complete, is at least impressively clear and voluminous. He is assumed also to have a well-organized and stable system of preferences, and a skill in computation that enables him to calculate, for the alternative courses of action that are available to him, which of these will permit him to reach the highest attainable point on his preference scale. (Simon, 1955, p. 99)

Both Simon and Lipman call for this picture to be revised, and this is what the proposals surveyed here all do. In order to account for vagueness in natural language in the context of these models, they peel away from this idealized picture and bring some of these assumptions down to earth. In the process, they point to ways in which we, as language learners and language users, are finite beings finding ways to cope with a highly complex and dynamic environment:

1. Our existence is temporally finite; language does not have an infinite amount of time to evolve, nor can it take an infinite time to be learned. The faster a language can start being useful, the better;

2. Language learning through experience has to rely on a limited number of observations. Not only is the state space typically much larger than one can survey in sufficient time, it is even potentially infinite and constantly changing;

3. A corollary of the former is that there will always be heterogeneity in a population of language learners, at the very least in their prior experience, since

each agent will have relied on a different set of observations. Furthermore, this information is not directly or fully accessible to others.

All of these observations support the weakening of modeling assumptions. The research surveyed here shows us some examples of assumptions which, when weakened, make vague signal use a natural outcome of certain evolutionary dynamics. But it gives us even more. It suggests ways in which the mechanisms that lead to vagueness can have positive effects that are extremely important in the context of the points just enumerated. We learned from O'Connor (2014b) and Franke and Correia (2018) that vague languages are quicker to converge and adapt, which is valuable given the finite and dynamic character of our experience (point 1). O'Connor (2017) also showed how generalization, an invaluable feature of any procedure for learning from a limited number of observations (point 2), leads to vagueness. We also learned from Franke and Correia (2018) that state confusability, a mechanism that leads to vague signal use, can have a homogenizing effect on vocabularies, potentially compensating for the heterogeneity of agents' experiences (point 3).

What do these observations tell us about rationality? GRL (O'Connor, 2014b) and the work of Franke and Correia (2018) both assume a picture of agents with a basic level of instrumental rationality, possibly limited awareness of the game, and a lack of strategic capabilities, adapting their behavior with only short-term gains in sight. These approaches introduce constraints on agent behavior or information processing that prevent the evolution of crisp signal use. But a crisp language would still have a higher expected utility than the evolved strategies. Agents in those models seem to be only as rational as the modelers allow them to be. Despite the plausibility of the mechanisms proposed (limited memory, imprecise calculation of expected utilities, generalization, state confusability), the results of these models feel somewhat bittersweet because of the hypothetical possibility of an ideal strategy, seemingly barred from the agents in an artificial manner. Couldn't a more rational agent evolve and drive the system into crisp signal use? Aren't we, human beings, that kind of agent?

Perhaps a deeper understanding of vagueness and the reasons for its pervasiveness in natural language are to be found only when we broaden the scope of the models employed. All the models discussed so far explore evolutionary dynamics for one homogeneous population playing one game. Different types of agents and different game setups are considered, but each of these different possible scenarios is tested separately. We see at least two ways in which one could embrace a more ecological perspective. The first is to think about meaning and vagueness from a more Wittgensteinian perspective. One can see each signaling game as embodying a particular language game. In Wittgenstein's picture of language, however, we do not play only one language game over the course of our existence; there is a plurality of them and which one an agent is engaged in at a particular moment is never clearly identified, neither are the exact benefits one might gain by choosing a certain behavior over another. These are furthermore not fixed in time; old language games fall out of fashion or stop being useful, and new ones emerge all the time (and in particular §23 Wittgenstein, 1953).

We can look for rationality at several levels in this pluralistic picture. First, as before, there is the actual behavior of a single agent in each actualized language game. As mentioned above in connection the soft-max choice function used by Franke, Jäger, and van Rooij (2011), behavior that strictly maximizes expected utility under uncertainty may be resource heavy, so it might be compatible with local strategic rationality that agents' production choices are stochastic. Second, if we look at behavior across many game types and contexts, there is also the level of an agent's internal theory of how words and phrases are likely to be used (or even normatively: how messages should be used), conditional on a given context. Notice that a single agent's rational beliefs about linguistic practices or linguistic meaning may well have to reflect the actual stochasticity: under natural assumptions about information loss, the best belief for prediction matches the actual distribution in the real world (Vehtari and Ojanen, 2012). In sum, both at the level of behavior and at the level of beliefs about use or meaning, we should expect to find vagueness. Still, despite the natural vagueness, there does not seem to be anything fundamentally missing or conceptually incoherent in a naturalistic, rationality- or optimality-driven explanation of what each agent is doing or what each agent believes about language, use and meaning.

Another way to go beyond locality is to work with more heterogeneous population models. The mechanisms that lead to vague signal use, as O'Connor (2014b) and Franke and Correia (2018) stress, have the aforementioned important advantages of faster speed of convergence, higher flexibility, and homogenization. The argument goes that these side-effects, by temporarily enabling a higher expected utility, could allow a population using some generalization (or affected by some imprecision) to take over. However, despite its intuitiveness, the argument is based on comparing isolated runs of different dynamics. The models do not allow the hypothesis to be tested, because they do not accommodate different populations evolving together.In the remainder of the paper we propose a way to do this based on the model of Franke and Correia (2018). We introduce two variants of a multi-population model of the imprecise imitation dynamics, where populations characterized by different imprecision values interact and evolve together. Using this model, we can better see under which conditions the hypothesized advantages of some imprecision can lead to the evolution of vagueness.

A.5 Two multi-population models of imprecise imitation

We build upon the model of Franke and Correia (2018), adding to the imprecise imitation dynamics support for multiple populations with different imprecision values evolving together. The resulting evolutionary dynamic has two layers: first, signaling strategies change by imprecise imitation following the model of Franke and Correia (2018); second, there is evolutionary selection at the level of the imprecision values themselves. More concretely, imagine agents who are born with varying perceptual abilities, subsequently learn a signaling strategy by imitation of other agents (either within or across populations), and depending on the success of the strategies they develop, are more or less likely to survive and reproduce. In such a setup, evolutionary processes will not only change the distribution of strategies within a population, where a population is identified by a shared level of imprecision in imitation, they will also promote those levels of imprecision which result in the development of more successful signal use.

This is closely related to the ideas behind theories of kin selection (W. D. Hamilton, 1964) and multi-level (or group) selection (Wilson, 1975). These theories build upon the hypothesis that selection acts not only to directly favor genes that result in behaviors that benefit individuals, but also to indirectly favor genes that lead to behaviors that benefit either genotypically (kin) or socially (group) related individuals (see Okasha, 2006, for more details). In our model there is a similar structure: there is a process of selection of behavior within each population, and levels of imprecision are selected across populations based on the strategies they give rise to. We believe there is an important difference with group and kin selection models in that the two selection processes in our model act on different entities: the inner shaping signaling behavior, and the outer selecting levels of imprecision. In any case, we intend our model to be descriptive rather than causal or explanatory. That means that we do not want to commit to seeing populations either as kin or as social groups, and use them as merely descriptive abstractions that allow us to capture the hypothetical impact of indirect selection processes.[6]

[6] We make this note because of the heated debate between the two theories. See Kohn (2008), and Kramer and Meunier (2016) for more details.

In the following, we lay down the formal details of our model. Let's start by defining A to be the set of imprecision values considered. For each value $\alpha \in A$, the proportion of its population is given by $P(\alpha)$, such that $\sum_{\alpha \in A} P(\alpha) = 1$. Each population has its own sender and receiver strategies, represented as σ^{α} and ρ^{α}. The probability that a given agent with imprecision α (or of type α) observes t_o when the actual state is t_a is given by $P_o^{\alpha}(t_o|t_a)$. If the same agent intends to realize t_i, the probability that she actually realizes t_r instead is given by $P_r^{\alpha}(t_r|t_i)$. Following Franke and Correia (2018, p. 26), we can then define the following values.

Probability that t_a is actual if t_o is observed by an agent of type α:

$$P_{\bar{o}}^{\alpha}(t_a|t_o) \propto P_a(t_a) P_o^{\alpha}(t_o|t_a)$$

Probability that a random sender of type α produces m when the actual state is t_a:

$$P_{\sigma}^{\alpha}(m|t_a) = \sum_{t_o} P_o^{\alpha}(t_o|t_a) \sigma^{\alpha}(m|t_o)$$

Probability that the actual state is t_a if a random sender of type α produced m:

$$P_{\bar{\sigma}}^{\alpha}(t_a|m) \propto P_a(t_a) P_{\sigma}^{\alpha}(m|t_a)$$

Probability that t_r is realized by a random receiver of type α in response to message m:

$$P_{\rho}^{\alpha}(t_r|m) = \sum_{t_i} P_r^{\alpha}(t_r|t_i) \rho^{\alpha}(t_i|m)$$

These formulations are merely parameterized versions of the single-population model. They encapsulate calculations that one can use to compute expected utilities and strategy update steps for each type. The latter, however, depend on the types of interaction that we imagine occurring between populations. In the following sections, we consider two different possibilities.

Tight population interaction

In a multi-population model with tight interaction between populations, each agent plays with, observes, and potentially imitates any other agent, regardless of their type. This has an impact on the expected utilities of sender and receiver strategies of each type, and on the update steps for those strategies. Let's start with the expected utilities. For a sender of type α, the expected utility of its strategy σ^{α} against all other receiver strategies $\rho^{\star}$, is given by:

$$\mathrm{EU}_{\sigma}^{\alpha}(m, t_o, \rho^{\star}) = \sum_{t_a} P_{\bar{o}}^{\alpha}(t_a|t_o) \sum_{\alpha' \in A} P(\alpha') \sum_{t_r} P_{\rho}^{\alpha'}(t_r|m) U(t_a, t_r)$$

and, for a receiver of type α, the expected utility of its strategy ρ^α against all other sender strategies $\sigma^\star$, is given by:

$$\mathrm{EU}^\alpha_\rho(t_i, m, \sigma^\star) = \sum_{\alpha' \in A} P(\alpha') \sum_{t_a} P^{\alpha'}_{\bar\sigma}(t_a|m) \sum_{t_r} P^\alpha_r(t_r|t_i) U(t_a, t_r)$$

Expected utilities thus take into account the existence of strategies of other types, and weigh the relevance of each type α' according to its relative proportion $P(\alpha')$.

Another important value to calculate has to do with the types that agents observe and imitate. In a model with tight interaction, we imagine this occurring across populations. Therefore, we can define the probability that a sender of type α observes a randomly sampled agent play message m for observed state t_o as:

$$P^\alpha_o(m|t_o) = \sum_{t_a} P^\alpha_{\bar o}(t_a|t_o) \sum_{\alpha' \in A} P(\alpha') P^{\alpha'}_\sigma(m|t_a)$$

and the probability that a receiver of type α observes a randomly sampled agent choose interpretation t_o given message m as:

$$P^\alpha_o(t_o|m) = \sum_{t_r} P^\alpha_o(t_o|t_r) \sum_{\alpha' \in A} P(\alpha') P^{\alpha'}_\rho(t_r|m)$$

Again, these calculations incorporate the probabilities that the imitating agent might observe the behavior of an agent of another type α', weighed by its relative proportion. Finally, the update step for a sender strategy of type α at time instant $i+1$ is given by:

$$\check\sigma^\alpha_{i+1}(m|t) \propto P^\alpha_o(m|t)\mathrm{EU}^\alpha_{\sigma_i}(m, t, \rho^\star_i)$$

and similarly for a receiver strategy of type α by:

$$\check\rho^\alpha_{i+1}(t|m) \propto P^\alpha_o(t|m)\mathrm{EU}^\alpha_{\rho_i}(t, m, \sigma^\star_i)$$

We here use $\check\sigma$ and $\check\rho$ since there is still an additional adjustment to these values to be calculated before we get the final strategies σ and ρ.

These formulations cover the evolution of the particular strategies of each type. We can think of this as the level of cultural evolution: agents are born with a certain level of imprecision and adopt strategies based on the behavior of others. Alongside this process, we can imagine another level of selection, where agents die and new agents are born. More successful agents have a higher likelihood of surviving and reproducing, giving rise to more agents with their level of imprecision. Levels of imprecision are thus subject to an evolutionary dynamic that is indirectly influenced by the cultural dynamic. Importantly, only the level of imprecision is passed on to new generations under this dynamic, not the actual strategies developed by the agents at the cultural level.

We model this process by changing the proportion of each type $P(\alpha)$ according to the replicator dynamic. A population of type α consists of agents employing both a sender and a receiver strategy, thus the overall fitness of the population must include the expected utilities of both. We could imagine this process happening at a different speed to the cultural process, in which case we could have different time scales. For the sake of simplicity, we choose to have them both happen at each time step, but calculate the changes occurring at this level of selection after the calculation for the other level. We define the proportion of type α at time step $i+1$ as:

$$P_{i+1}(\alpha) \propto P_i(\alpha)(\mathrm{EU}^{\alpha}_{\sigma_{i+1}}(m,t,\rho^{\star}_i) + \mathrm{EU}^{\alpha}_{\rho_{i+1}}(t,m,\sigma^{\star}_i))$$

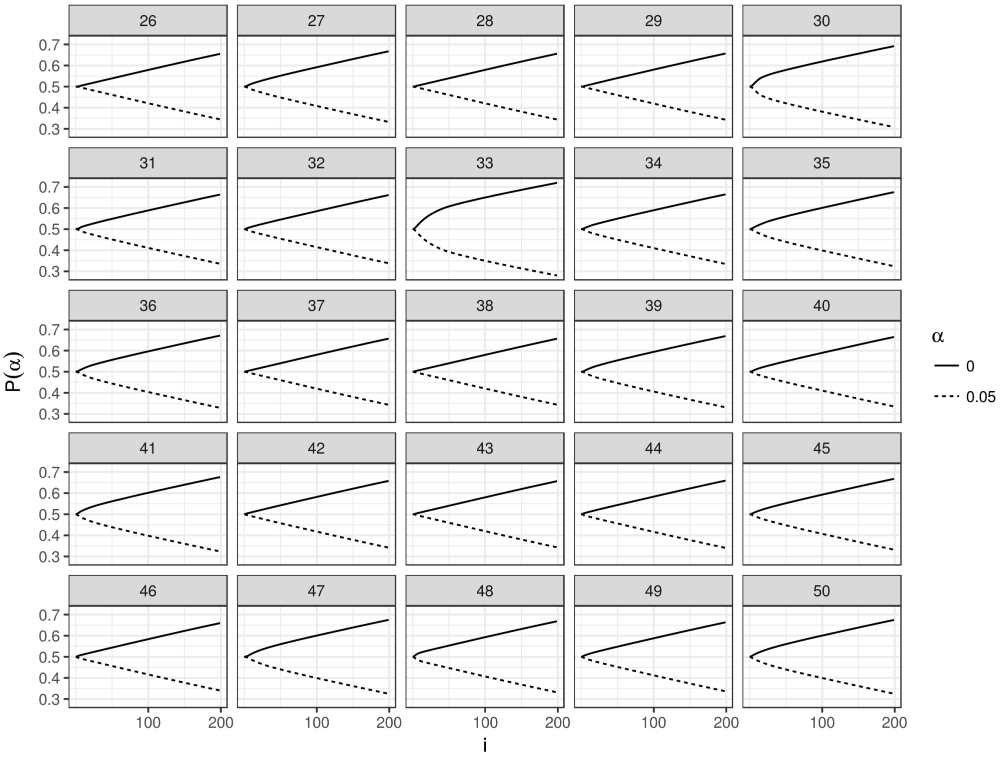
In order to additionally account for the fact that strategies are not passed on to new generations, we mix the evolved strategies of a certain type σ^α and ρ^α with new random strategies $\tilde{\sigma}^\alpha$ and $\tilde{\rho}^\alpha$ (generated at each time step). The idea is that, at each time step, a certain percentage of each population will consist of "newborn" agents, *i.e.* agents that haven't yet had time to evolve their strategies. We define a parameter γ that quantifies this percentage, or as we can also call it the birth rate, which we consider to be the same for every population. This mixing finally defines the strategies for time step $i+1$ and can be described in the following formulas:

$$\sigma^{\alpha}_{i+1}(m|t) = (1-\gamma)\check{\sigma}^{\alpha}_{i+1}(m|t) + \gamma\tilde{\sigma}^{\alpha}_{i+1}(m|t)$$
$$\rho^{\alpha}_{i+1}(t|m) = (1-\gamma)\check{\rho}^{\alpha}_{i+1}(t|m) + \gamma\tilde{\rho}^{\alpha}_{i+1}(t|m)$$

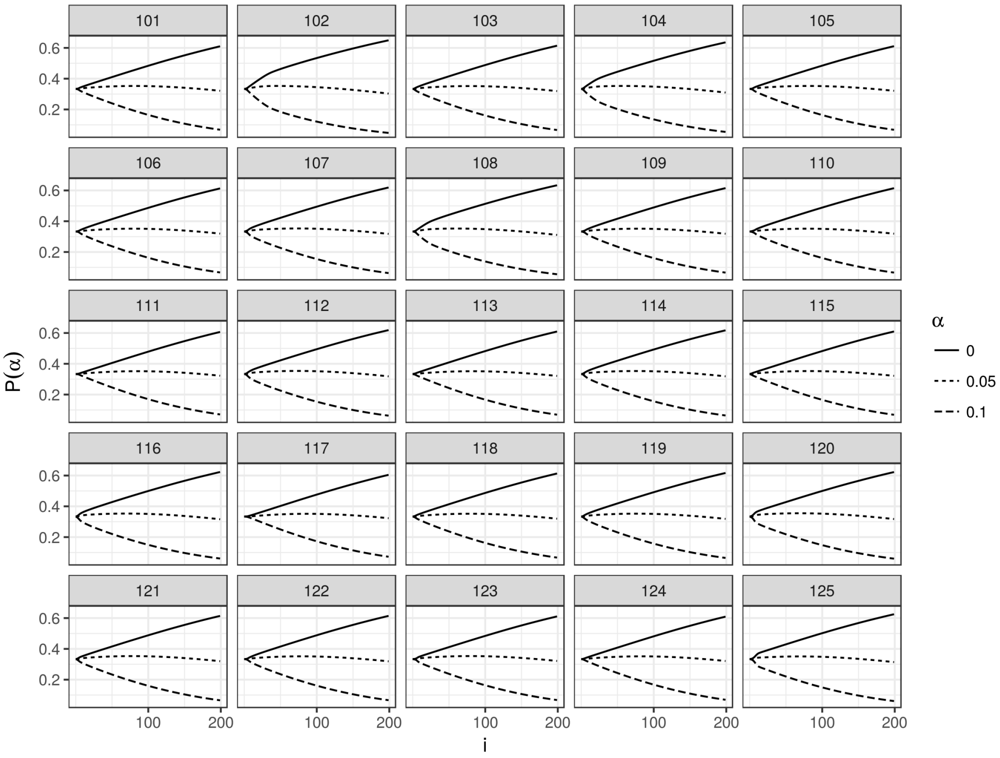
Simulation results We performed 25 simulation runs of this model for each of three population scenarios: only one population with $\alpha = 0$ (for reference), two populations with $\alpha \in \{0, 0.05\}$, and three populations with $\alpha \in \{0, 0.05, 0.1\}$. For each scenario, starting proportions $P(\alpha)$ were equal for each value of α. Given the observations by Franke and Correia (2018) that state space size and tolerance parameter β do not result in important qualitative difference, we fixed these values at $n_S = 30$ and $\beta = 0.1$. We also used a fixed uniform distribution for the priors and a message space with 2 messages. Each type started with its own randomly generated strategy. Regarding the duration of the simulations, due to the mixing in of new individuals into the population (birth rate was fixed at $\gamma = 0.05$), the convergence criteria is no longer applicable because each strategy is randomly perturbed at each time step. Therefore, all simulation runs were stopped after 200 iterations.

The first thing to observe from the simulation results is that the population with no imprecision ($\alpha = 0$) dominated the other populations in every run.

In Figure A.2 we plot the evolution of the proportion of each population in the two-population and three-population scenarios for all trials. As the plot shows, the



(a) Two-population scenario.



(b) Three-population scenario.

Figure A.2: Evolution of population proportions through time for each simulation trial of the tight interaction model. Numbers on top of each plot identify each trial.

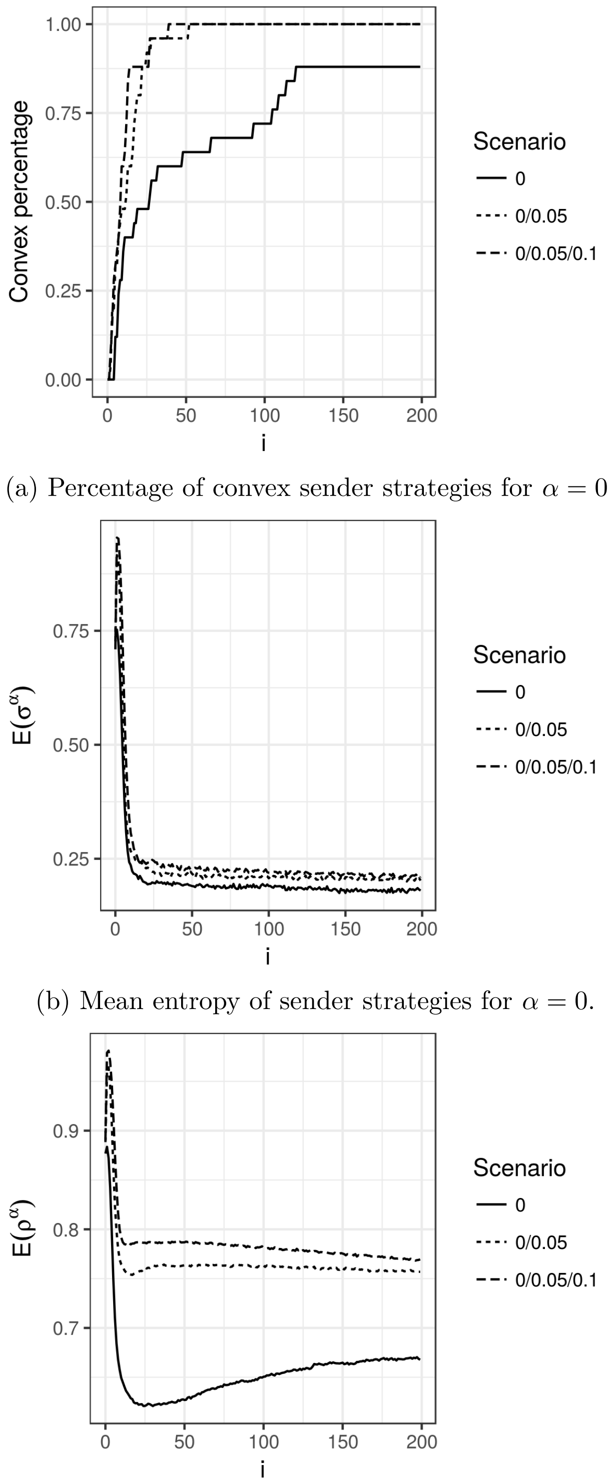
population with no imprecision steadily increased its proportion against the others in every trial. In the three-population scenario, the proportion of type $\alpha = 0.05$ sees a slight increase in the beginning of the simulation, but inexorably starts a downward trend. These observations go against the expectation of Franke and Correia (2018) that faster convergence to a convex strategy by populations with a certain level of imprecision could give them a temporary advantage to take over and eliminate other types. The reason for this is interesting in itself. What happens is that, because of the tight interaction between populations, the strategies of each type evolve in close tandem with each other. One of the consequences of this is that the population with no imprecision reaches convexity faster than it would on its own because of the interaction with the populations with imprecision. We can see this effect by looking into the percentage of trials with convex sender strategies at a given iteration, for each scenario, and comparing the three scenarios: populations with no imprecision evolving alone, two populations ($\alpha \in \{0, 0.05\}$), and three populations ($\alpha \in \{0, 0.05, 0.1\}$).

This is plotted in Figure A.3a. What we see is more trials reaching convexity earlier for the multi-population scenarios when compared with the single-population scenario. This effect precludes the hypothesized temporary advantage of imprecision manifesting itself, but it can be seen as a positive influence of the population(s) with imprecision on the population with no imprecision.

The flip side of this tight connection between populations is that strategies evolved by populations with no imprecision are also more vague (in the sense defined in section A.3). This is visible by looking at mean entropy values, namely sender strategy entropy, plotted in Figure A.3b, and receiver strategy entropy, plotted in Figure A.3c. The values for the population with no imprecision are clearly higher in the scenarios where it evolves together with populations with imprecision than in the scenario where it evolves on its own. Given the trends in population proportions, one expects this to eventually be eliminated when the population with no imprecision finally takes over the others, but it is interesting to observe that while populations with vague strategies persist, the population with no imprecision takes much longer to evolve a fully crisp strategy.

Loose population interaction

We can also imagine a scenario where populations interact more loosely. Namely, the dynamic we want to model here is one where agents of a certain type α imitate and learn only from other agents of the same type, but nevertheless use their signaling strategies with agents from all types. In order to capture this, we need



(a) Percentage of convex sender strategies for $\alpha = 0$.

(b) Mean entropy of sender strategies for $\alpha = 0$.

(c) Mean entropy of receiver strategies for $\alpha = 0$.

Figure A.3: Development of some metrics through time for the $\alpha = 0$ population in each of three scenarios: evolving alone ('0'), with an $\alpha = 0.05$ population ('0/0.05'), and with an additional $\alpha = 0.1$ population ('0/0.05/0.1').

to make changes to some calculations. Any formula that is not redefined in this section should be assumed to remain the same. First, if agents learn only within their population, the imitation dynamic needs to consider only the expected utility against agents of that population. We thus define the following expected utilities for a sender of type α:

$$\mathrm{EU}^{\alpha}_{\sigma}(m, t_o, \rho^{\alpha}) = \sum_{t_a} P^{\alpha}_{\bar{o}}(t_a|t_o) \sum_{t_r} P^{\alpha}_{\rho}(t_r|m) U(t_a, t_r)$$

and for a receiver of type α:

$$\mathrm{EU}^{\alpha}_{\rho}(t_i, m, \sigma^{\alpha}) = \sum_{t_a} P^{\alpha}_{\bar{\sigma}}(t_a|m) \sum_{t_r} P^{\alpha}_{r}(t_r|t_i) U(t_a, t_r)$$

The main difference from the previous model is that these expected utilities are not calculated against all populations ($\sigma^{\star}$ and $\rho^{\star}$) but only against the agent's own type (σ^{α} and ρ^{α}). This also implies that population proportions do not play a role at this level of selection.

Regarding the imitation process, we can redefine the probability that a sender of type α observes a randomly sampled agent play message m for observed state t_o as:

$$P^{\alpha}_{o}(m|t_o) = \sum_{t_a} P^{\alpha}_{\bar{o}}(t_a|t_o) P^{\alpha}_{\sigma}(m|t_a)$$

and the probability that a receiver of type α observes a randomly sampled agent choose interpretation t_o given message m as:

$$P^{\alpha}_{o}(t_o|m) = \sum_{t_r} P^{\alpha}_{o}(t_o|t_r) P^{\alpha}_{\rho}(t_r|m)$$

Again, the main difference is that agents make observations within their own population, so to model a randomly sampled agent one needs only to take into account agents of the same type. Population proportions again do not play a role in these calculations. Based on these formulas, we can define the update step for a sender strategy of type α at time instant $i + 1$ as:

$$\check{\sigma}^{\alpha}_{i+1}(m|t) \propto P^{\alpha}_{o}(m|t) \mathrm{EU}^{\alpha}_{\sigma_i}(m, t, \rho^{\alpha}_{i})$$

and similarly for a receiver strategy of type α as:

$$\check{\rho}^{\alpha}_{i+1}(t|m) \propto P^{\alpha}_{o}(t|m) \mathrm{EU}^{\alpha}_{\rho_i}(t, m, \sigma^{\alpha}_{i})$$

Note that, because imitation and learning occur only within populations, these formulations are essentially the same as for the single-population model of Franke and Correia (2018), only parameterized by type α.

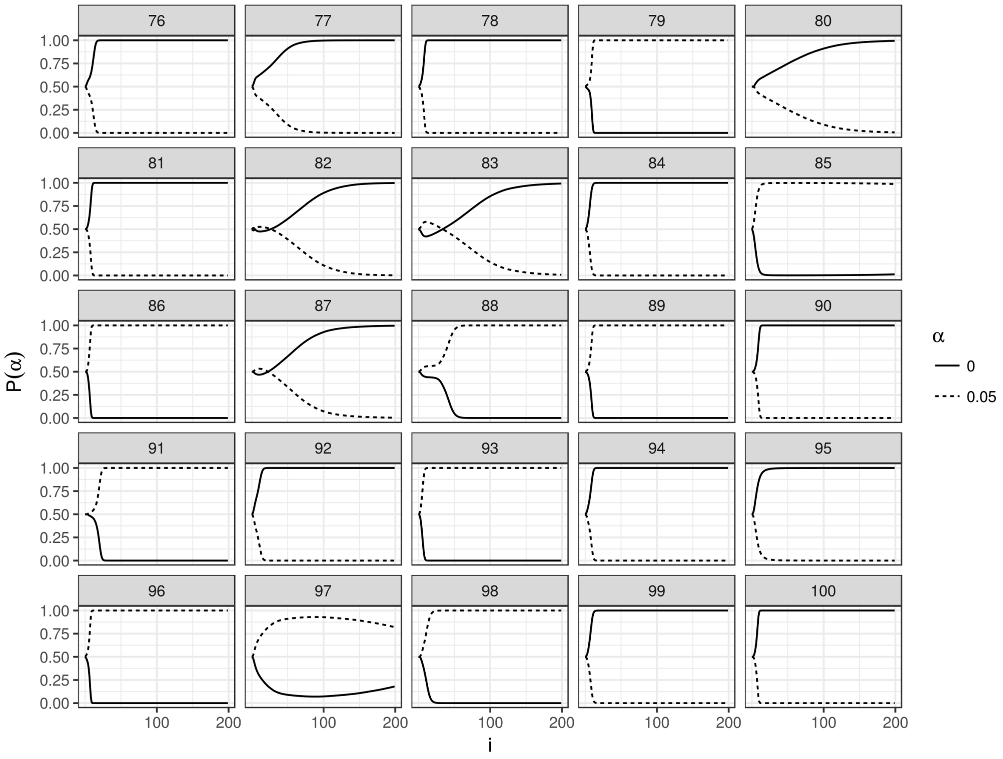
The selection process between different populations still follows the same motivation as before: agents of a certain type that evolve successful strategies (with respect to other types) will be more likely to survive and reproduce, benefiting the proportion of their population. The formulation of the dynamic for the proportion of each type $P(\alpha)$ thus stays the same. In order to avoid confusion, we want to stress that this means that these calculations rely on the definitions of expected utility across populations (*i.e.* $\mathrm{EU}^{\alpha}_{\sigma}(m, t_o, \rho^{\star})$ and $\mathrm{EU}^{\alpha}_{\rho}(t_i, m, \sigma^{\star})$) and not the newly introduced $\mathrm{EU}^{\alpha}_{\sigma}(m, t_o, \rho^{\alpha})$ and $\mathrm{EU}^{\alpha}_{\rho}(t_i, m, \sigma^{\alpha})$.

Simulation results We ran the same number of simulation trials under the same conditions as for the model with tight population interaction.

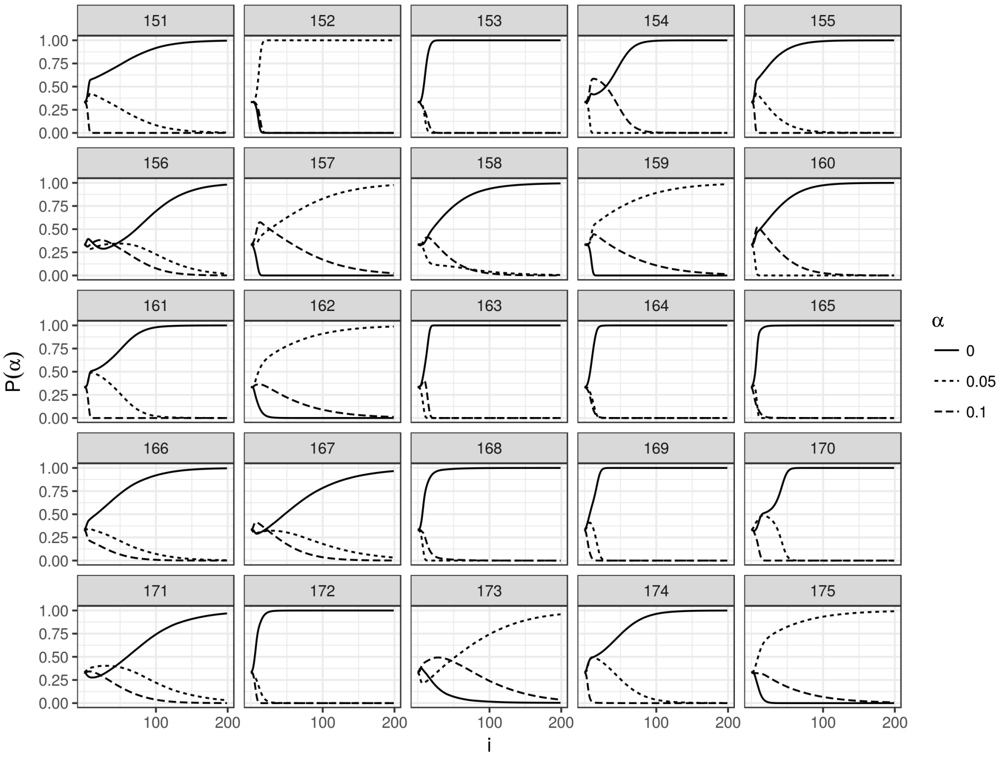
In Figure A.4 we plot the evolution of population proportions for all trials of the multi-population scenarios. The first thing to observe is that proportions evolve faster than in the model with tight interaction. Whereas in the latter no given population ever reached much more than 70% proportion, in this model we see that many trials resulted in one population fully dominating the others. In the two-population scenario, some population reached at least 99% in 24 out of 25 trials. In the three-population scenario, this happened in 18 out of 25 trials. More interestingly for our investigation, some trials actually resulted in the population with $\alpha = 0$ being dominated. For the two-population scenario, $\alpha = 0.05$ fully reached 100% in 8 trials (79, 86, 88, 89, 91, 93, 96, 98), a point from which it is technically impossible for the other population to recover. In most cases the dominating population gains its ground from the start, but in 5 cases we see a temporary advantage of $\alpha = 0.05$ that is then lost to the other population. In one interesting trial (85), $\alpha = 0.05$ reached 99.87% only to then steadily start losing ground to $\alpha = 0$. For the three-population scenario, $\alpha = 0$ was reduced to 0% in 4 trials (152, 157, 159, and 175). In all of those cases, $\alpha = 0.05$ clearly has the upper hand over $\alpha = 0.1$, despite a temporary advantage of the latter in some trials.

Because of the loose interaction between populations, different types can now evolve separately. One consequence of this is that populations with a certain level of imprecision again reach convexity faster than those without imprecision.

In Figure A.5 we plot, for each trial, the first iteration when each type reached convexity. What we see is that, even though $\alpha = 0$ usually reaches convexity later, this is not always the case. This has certainly to do with the initial conditions of each trial, since the randomly generated strategies can simply by chance be more favourable to reaching convexity. More importantly, we also see that reaching convexity sooner is not a sufficient condition for achieving population dominance. There

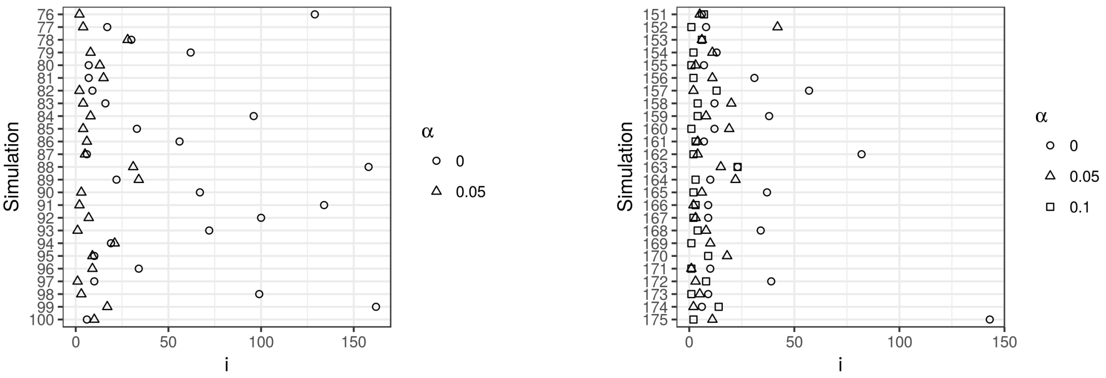


(a) Two-population scenario.



(b) Three-population scenario.

Figure A.4: Evolution of population proportions through time for each simulation trial of the loose interaction model. Numbers on top of each plot identify each trial.



(a) Two populations scenario.

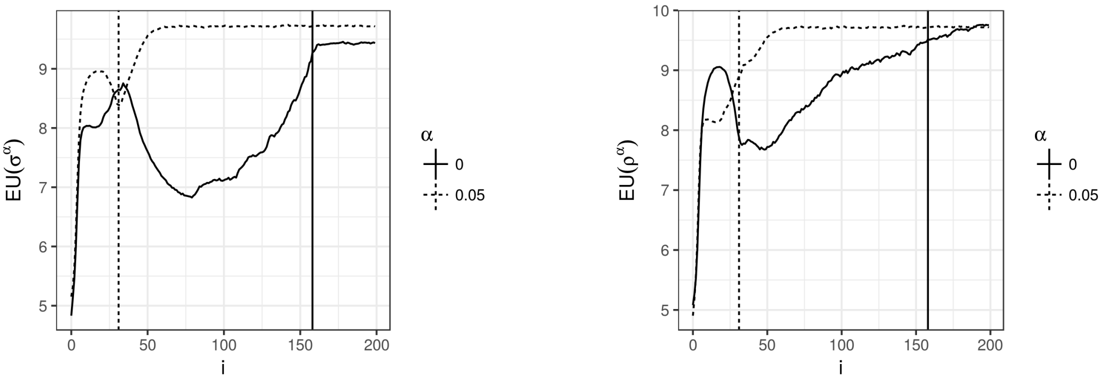
(b) Three populations scenario.

Figure A.5: First iteration with convex sender strategies for each simulation trial of the loose interaction model.

were many trials where another type reached convexity much sooner than $\alpha = 0$ but its population was completely dominated nevertheless (*e.g.* 76, 84, 92, and 99 in the two-population scenario; 156, 165, 168, and 172 in the three-population scenario). It also does not seem to be fully necessary, given a few examples where $\alpha = 0$ reached convexity early on and another type ended up dominating (89 in the two-population scenario and 152 in the three-population scenario).

Another consequence of the populations evolving separately in this model is that they do not necessarily evolve towards the same equilibrium. In a sim-max game such as the one set up here, there are only two stable equilibria. These are the two Voronoi languages of the kind shown in Figure A.1: one where the first message is used for the first half of the state space, and the other where the second message is used for this region. In the multi-population model with loose interaction, each population evolves independently towards one of these two equilibria. An important factor determining which language strategies converge to is the random initial population configuration. The populations are, however, not fully independent, since the process of selection of the level of precision relies on the expected utility of one population playing against itself and the others. And this is important because strategies in one equilibrium get the lowest payoff possible against strategies in the other equilibrium. Now, if two populations evolve towards different equilibria, whatever advantage one population has playing against itself could trigger a runaway effect by causing an increase in its proportion, which in turn will increase the population's relative expected utility, potentially increasing their proportion further in the next round, and so forth. In this case, one would expect that faster convergence towards convexity would be especially important for a population's success.

In the two-population scenario, of the 8 trials where $\alpha = 0$ was reduced to 0%, all



(a) Sender expected utility.

(b) Receiver expected utility.

Figure A.6: Expected utilities for trial 88 of the loose interaction model. Vertical lines demarcate, for each type, the iteration where convexity was achieved.

but two (88, 98) ended with the two populations each close to different equilibria. In the three-population scenario, the population with $\alpha = 0$ evolved towards the same equilibrium as the other two in 1 of the trials where it was eliminated. In the other 3 trials, in 2 it evolved to a different equilibrium than the other two, and in 1 it evolved towards the same equilibrium as $\alpha = 0.05$ (but different to the one $\alpha = 0.1$ evolved towards). Again, we do not find a clear case for this being a decisive factor in the success of populations with some imprecision. The same goes for the impact of an initial advantage in expected utility creating the runaway effect we just mentioned. Despite none of these three factors (including reaching convexity sooner) sentencing the demise of $\alpha = 0$ with certainty, they do seem to conspire together to bring it about. In most trials where the type was eliminated, in both the two- and three-population scenarios, $\alpha = 0$ ended up evolving towards a different equilibrium than the other types, and either had an initial disadvantage in expected utility, or reached convexity much later. These factors can thus be seen as indicators at best, and the story of how a population with imprecision ends up dominating one without it seems to be more complicated than expected.

This is not surprising, since we are facing a complex dynamical system with various interacting components (sender and receiver strategies and multiple populations).

Just as an example of this, in Figure A.6 we plot the expected utilities of sender and receiver strategies for both populations of trial 88. The curves up to where $\alpha = 0.05$ achieved convexity seem to suggest that the sender strategy of that type was evolving towards the same equilibrium as the receiver strategy of the other. The moment where the population reaches convexity marks the point where the sender strategy of $\alpha = 0.05$ aligns with its type's receiver strategy and this coincides with the moment when the population seems to gain real traction over the other type (see again the plot for trial 88 in Figure A.5). In this particular case, reaching convexity seems to have made an important difference.

A.6 Conclusions

Vagueness presents a challenge to both procedural and instrumental pictures of rationality alike. In the context of game-theoretical models of language, this takes the form of a question about evolution: how can vagueness persist if it characterizes demonstrably less efficient communication? Most existing proposals in the literature attempt to explain this by considering agents with some degree of bounded rationality. Despite the many ways in which our rationality is inevitably limited, the argument for the pervasiveness of vagueness in natural language would be much stronger if one could also find associated advantages. In this paper, we argued that finding those might require us to go beyond a local notion of rationality, as two of the proposals reviewed here (O'Connor, 2014b; Franke and Correia, 2018) suggest. We advocate moving towards an ecological approach, studying vagueness in more heterogeneous ecosystems. Language is part of a very complex system (Beckner et al., 2009) that involves many components interacting in often unpredictable ways. An ecology of vagueness would involve studying models where different populations can evolve and interact with each other, where different language games can be played between individuals, where the environment is uncertain and changing, or anything else that more closely approximates the real context of language evolution.

In light of this picture, we have proposed two variants of a concrete multi-population signaling model to test the hypothesis (Franke and Correia, 2018) that certain features of imprecise imitation, like promoting faster convergence and regularity, could prove beneficial in contrast with full precision. Analysing these models did not provide us with a clear-cut answer, and revealed that the story is much more nuanced than initially expected. In a variant where populations with and without imprecision interact tightly, although precision always has the upper hand, vagueness seems to take a long time to be weeded out. When we let populations interact more loosely, we see a more complex pattern of outcomes. These include scenarios where imprecise imitation dominates over full precision, showing that strategies with vagueness can actually, under certain circumstances, be more successful. Bringing several populations together in a more complex ecosystem thus allowed us to not only spell out and test the original intuition, but also learn about unforeseen effects.

These models thus serve as an example of how moving to a more global perspective on rationality can allow us to achieve a more detailed awareness of the complex interactions that might be involved in sustaining vagueness in natural language. They can be seen, we believe, as a first step towards an ecology of vagueness.

Bibliography

Akmajian, Adrian, Ann Kathleen Farmer, Lee Bickmore, Richard A. Demers, and Robert M. Harnish, eds. (2017). *Linguistics: An Introduction to Language and Communication.* Seventh edition. Cambridge: MIT Press.

Alexander, J. McKenzie (2014). “Learning to Signal in a Dynamic World”. In: *The British Journal for the Philosophy of Science* 65.4, pp. 797–820.

Alexander, J. McKenzie, Brian Skyrms, and Sandy L. Zabell (2012). “Inventing new signals”. In: *Dynamic Games and Applications* 2.1, pp. 129–145.

Allen, Colin and Jacob Neal (2019). “Teleological Notions in Biology”. In: *The Stanford Encyclopedia of Philosophy.* Ed. by Edward N. Zalta. Spring 2019. Metaphysics Research Lab, Stanford University.

Andrews, Michael, Edward Thommes, and Monica G. Cojocaru (2015). “Replicator Dynamics of Axelrod’s Norms Games”. In: *Interdisciplinary Topics in Applied Mathematics, Modeling and Computational Science.* Ed. by Monica G. Cojocaru, Ilias S. Kotsireas, Roman N. Makarov, Roderick V. N. Melnik, and Hasan Shodiev. Vol. 117. Springer Proceedings in Mathematics & Statistics. Cham: Springer, pp. 29–34.

Argiento, Raffaele, Robin Pemantle, Brian Skyrms, and Stanislav Volkov (2009). “Learning to Signal: Analysis of a Micro-level Reinforcement Model”. In: *Stochastic Processes and their Applications* 119.2, pp. 373–390.

Augustine (1993). “The Confessions”. In: *Saint Augustine.* Ed. by Mortimer J. Adler, Clifton Fadiman, and Philip W. Goetz. Trans. by R.S. Pine-Coffin. Fourth printing. Great Books of the Western World 16. Encyclopedia Britannica, Inc., pp. 1–159.

Austin, J. L. (1962). *How to Do Things with Words.* William James Lectures. Cambridge: Harvard University Press.

Axelrod, Robert (1986). “An Evolutionary Approach to Norms”. In: *American Political Science Review* 80.04, pp. 1095–1111.

Baker, Gordon (2004). *Wittgenstein’s Method: Neglected Aspects.* John Wiley & Sons, Inc.

Baker, Gordon and P. M. S. Hacker (1980). *Wittgenstein: Understanding and Meaning.* An Analytical Commentary on the Philosophical Investigations 1. Oxford: Blackwell.

— (1985). *Wittgenstein: Rules, Grammar and Necessity.* An Analytical Commentary on the Philosophical Investigations 2. Oxford: Blackwell.

Barrett, Jeffrey A. (2006). "Numerical Simulations of the Lewis Signaling Game: Learning Strategies, Pooling Equilibria, and the Evolution of Grammar". In: *Institute for Mathematical Behavioral Sciences.*

— (2007). "Dynamic Partitioning and the Conventionality of Kinds". In: *Philosophy of Science* 74.4, pp. 527–546.

— (2009). "The Evolution of Coding in Signaling Games". In: *Theory and Decision* 67.2, pp. 223–237.

— (2010). "Faithful Description and the Incommensurability of Evolved Languages". In: *Philosophical Studies* 147.1, pp. 123–137.

— (2012). "On the Coevolution of Basic Arithmetic Language and Knowledge". In: *Erkenntnis* 78.5, pp. 1025–1036.

— (2013a). "On the Coevolution of Theory and Language and the Nature of Successful Inquiry". In: *Erkenntnis* 79.4, pp. 821–834.

— (2013b). "The Evolution of Simple Rule-Following". In: *Biological Theory* 8.2, pp. 142–150.

Barrett, Jeffrey A. and Kevin J. S. Zollman (2009). "The Role of Forgetting in the Evolution and Learning of Language". In: *Journal of Experimental & Theoretical Artificial Intelligence* 21.4, pp. 293–309.

Beaney, Michael (2016). "Analysis". In: *The Stanford Encyclopedia of Philosophy.* Ed. by Edward N. Zalta. Summer 2016. Metaphysics Research Lab, Stanford University.

Beckner, Clay, Richard Blythe, Joan Bybee, Morten H. Christiansen, William Croft, Nick C. Ellis, John Holland, Jinyun Ke, Diane Larsen-Freeman, and Tom Schoenemann (2009). "Language Is a Complex Adaptive System: Position Paper". In: *Language Learning* 59, pp. 1–26.

Beggs, Alan W. (2005). "On the Convergence of Reinforcement Learning". In: *Journal of Economic Theory* 122.1, pp. 1–36.

Ben-Menahem, Yemima (1998). "Explanation and Description: Wittgenstein on Convention". In: *Synthese* 115.1, pp. 99–130.

Bennett, Peter G. (1977). "Toward a Theory of Hypergames". In: *Omega* 5.6, pp. 749–751.

Benz, Anton, Gerhard Jäger, and Robert van Rooij, eds. (2006). *Game Theory and Pragmatics*. Palgrave Studies in Pragmatics, Language and Cognition. London: Palgrave Macmillan.

Björnerstedt, Jonas and Jörgen Weibull (1995). "Nash Equilibrium and Evolution by Imitation". In: *The Rational Foundations of Economic Behavior*. Ed. by Kenneth Arrow, Christian Schmidt, Mark Perlman, and Enrico Colombatto. London: Macmillan, pp. 155–171.

Blackburn, Simon (2006). *Truth: A Guide for the Perplexed*. London: Penguin Books.

Börgers, Tilman and Rajiv Sarin (1997). "Learning Through Reinforcement and Replicator Dynamics". In: *Journal of Economic Theory* 77.1, pp. 1–14.

Boyd, Robert and Peter J. Richerson (1985). *Culture and the Evolutionary Process*. Chicago: University of Chicago Press.

Brochhagen, Thomas (2017). "Signalling under Uncertainty: Interpretative Alignment without a Common Prior". In: *The British Journal for the Philosophy of Science*.

Brown, George W. (1951). "Iterative Solutions of Games by Fictitious Play". In: *Activity Analysis of Production and Allocation*. Ed. by T.C. Koopmans. New York: John Wiley & Sons, Inc.

Brown, James Robert (1991). *The Laboratory of the Mind: Thought Experiments in the Natural Sciences*. Philosophical Issues in Science. London; New York: Routledge.

Brown, James Robert and Yiftach Fehige (2017). "Thought Experiments". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Summer 2017. Metaphysics Research Lab, Stanford University.

Catteeuw, David and Bernard Manderick (2014). "The Limits and Robustness of Reinforcement Learning in Lewis Signalling Games". In: *Connection Science* 26.2, pp. 161–177.

Connelly, Brian L., S. Trevis Certo, R. Duane Ireland, and Christopher R. Reutzel (2011). "Signaling Theory: A Review and Assessment". In: *Journal of Management* 37.1, pp. 39–67.

Cooper, Rachel (2005). "Thought Experiments". In: *Metaphilosophy* 36.3, pp. 328–347.

Correia, José Pedro (2013). "The Bivalent Trap: Vagueness, Theories of Meaning, and Identity". MA thesis. Universiteit van Amsterdam.

— (2019). "Analysis and Explanation in the *Philosophical Investigations*". In: *Logical Analysis and History of Philosophy*. To appear.

Correia, José Pedro and Michael Franke (2019). “Towards an Ecology of Vagueness”. In: *Vagueness and Rationality in Language Use and Cognition.* Ed. by Richard Dietz. Language, Cognition, and Mind. Cham: Springer, pp. 87–113.

Correia, José Pedro and Radek Ocelák (2019). “Towards More Realistic Modeling of Linguistic Color Categorization”. In: *Open Philosophy* 2.1, pp. 160–189.

Cover, Thomas M. and Joy A. Thomas (2006). *Elements of Information Theory.* Second edition. Hoboken: Wiley-Blackwell.

Crawford, Vincent P. and Joel Sobel (1982). “Strategic Information Transmission”. In: *Econometrica* 50.6, pp. 1431–1451.

Davis, Isaac (2017). “Understanding the Role of Perception in the Evolution of Human Language”. In: *Proceedings of the 39th Annual Meeting of the Cognitive Science Society.* Austin: Cognitive Science Society, Inc., pp. 1902–1907.

Dawkins, Richard (1976). *The Selfish Gene.* Oxford University Press.

Di Paolo, Ezequiel A., Jason Noble, and Seth Bullock (2000). “Simulation Models as Opaque Thought Experiments”. In: *Proceedings of the Seventh International Conference on Artificial Life.* Ed. by Mark A. Bedau, John S. McCaskill, Norman H. Packard, and Steen Rasmussen. Cambridge: MIT Press, pp. 497–506.

Dummett, Michael (1975). “What is a Theory of Meaning?” In: *Mind and Language.* Ed. by Samuel Guttenplan. Oxford: Oxford University Press, pp. 97–138.

— (1996). “Frege’s Myth of the Third Realm”. In: *Frege and Other Philosophers.* Oxford University Press, pp. 249–262.

El Skaf, Rawad and Cyrille Imbert (2013). “Unfolding in the Empirical Sciences: Experiments, Thought Experiments and Computer Simulations”. In: *Synthese* 190.16, pp. 3451–3474.

Fawcett, Tim W., Steven Hamblin, and Luc-Alain Giraldeau (2013). “Exposing the Behavioral Gambit: the Evolution of Learning and Decision Rules”. In: *Behavioral Ecology* 24.1, pp. 2–11.

Fogelin, Robert J. (1976). *Wittgenstein.* London: Routledge & Kegan Paul.

Franke, Michael (2014a). “Creative Compositionality From Reinforcement Learning in Signaling Games”. In: *The Evolution of Language: Proceedings of the 10th International Conference.* Ed. by Erica A. Cartmill, Seán Roberts, Heidi Lyn, and Hannah Cornish. World Scientific, pp. 82–89.

— (2014b). “Pragmatic Reasoning About Unawareness”. In: *Erkenntnis* 79.4, pp. 729–767.

— (2017). “Game Theory in Pragmatics: Evolution, Rationality, and Reasoning”. In: *Oxford Research Encyclopedia of Linguistics.*

Franke, Michael and José Pedro Correia (2018). “Vagueness and Imprecise Imitation in Signalling Games”. In: *The British Journal for the Philosophy of Science* 69.4, pp. 1037–1067.

Franke, Michael, Gerhard Jäger, and Robert van Rooij (2011). “Vagueness, Signaling and Bounded Rationality”. In: *New Frontiers in Artificial Intelligence*. Berlin, Heidelberg: Springer, pp. 45–59.

Franke, Michael and Elliott Wagner (2014). “Game Theory and the Evolution of Meaning”. In: *Language and Linguistics Compass* 8.9, pp. 359–372.

Frigg, Roman and Stephan Hartmann (2018). “Models in Science”. In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Summer 2018. Metaphysics Research Lab, Stanford University.

Galeazzi, Paolo and Michael Franke (2017). “Smart Representations: Rationality and Evolution in a Richer Environment”. In: *Philosophy of Science* 84.3, pp. 544–573.

Gilboa, Itzhak and Akihiko Matsui (1991). “Social Stability and Equilibrium”. In: *Econometrica* 59.3, pp. 859–867.

Glock, Hans-Johann (2007). “Perspectives on Wittgenstein: An Intermittently Opinionated Survey”. In: *Wittgenstein and His Interpreters: Essays in Memory of Gordon Baker*. Ed. by Guy Kahane, Edward Kanterian, and Oskari Kuusela. Reprint edition. Chichester: Wiley-Blackwell, pp. 37–65.

Godfrey-Smith, Peter and Manolo Martínez (2013). “Communication and Common Interest”. In: *PLOS Computational Biology* 9.11, pp. 1–6.

Grafen, Alan (1990). “Biological Signals as Handicaps”. In: *Journal of Theoretical Biology* 144.4, pp. 517–546.

Grice, H. P. (1957). “Meaning”. In: *The Philosophical Review* 66.3, pp. 377–388.

Grim, Patrick, Trina Kokalis, Ali Alai-Tafti, Nicholas Kilb, and Paul St Denis (2004). “Making Meaning Happen”. In: *Journal of Experimental & Theoretical Artificial Intelligence* 16.4, pp. 209–243.

Gruender, David (1962). “Wittgenstein on Explanation and Description”. In: *The Journal of Philosophy* 59.19, pp. 523–530.

Guala, Francesco (2013). “The Normativity of Lewis Conventions”. In: *Synthese* 190.15, pp. 3107–3122.

Haack, Robin (1982). “Wittgenstein’s Pragmatism”. In: *American Philosophical Quarterly* 19.2, pp. 163–171.

Haack, Susan (1997). “Vulgar Rortyism: Review of *Pragmatism: A Reader* by Louis Menand”. In: *The New Criterion*.

Hacker, P. M. S. (1990). *Wittgenstein: Meaning and Mind*. An Analytical Commentary on the Philosophical Investigations 3. Oxford: Blackwell.

Hacker, P. M. S. (1996). *Wittgenstein: Mind and Will.* An Analytical Commentary on the Philosophical Investigations 4. Oxford: Blackwell.

— (2007). “Gordon Baker’s Late Interpretation of Wittgenstein”. In: *Wittgenstein and His Interpreters: Essays in Memory of Gordon Baker.* Ed. by Guy Kahane, Edward Kanterian, and Oskari Kuusela. Reprint edition. Chichester: Wiley-Blackwell, pp. 88–122.

— (2012). “Wittgenstein on Grammar, Theses and Dogmatism”. In: *Philosophical Investigations* 35.1, pp. 1–17.

Hagen, Edward H., Nick Chater, Charles R. Gallistel, Alasdair Houston, Alex Kacelnik, Tobias Kalenscher, Daniel Nettle, Danny Oppenheimer, and David W. Stephens (2012). “Decision Making: What Can Evolution Do for Us?” In: *Evolution and the Mechanisms of Decision Making.* Ed. by Peter Hammerstein and Jeffrey R. Stevens. Cambridge: MIT Press.

Hale, Bob and Crispin Wright, eds. (1997). *A Companion to the Philosophy of Language.* Blackwell Companions to Philosophy. Oxford: Blackwell Publishers.

Hamilton, Nicholas E. and Michael Ferry (2018). “ggtern: Ternary Diagrams Using ggplot2”. In: *Journal of Statistical Software, Code Snippets* 87.3, pp. 1–17.

Hamilton, William D. (1964). “The Genetical Evolution of Social Behaviour. II”. In: *Journal of Theoretical Biology* 7.1, pp. 17–52.

Hattiangadi, Anandi (2007). *Oughts and Thoughts: Rule-Following and the Normativity of Content.* Oxford: Oxford University Press.

Hauser, Marc D., Noam Chomsky, and William Tecumseh Fitch (2002). “The Faculty of Language: What is It, Who Has It, and How Did It Evolve?” In: *Science* 298.5598, pp. 1569–1579.

Hofbauer, Josef and Simon M. Huttegger (2008). “Feasibility of Communication in Binary Signaling Games”. In: *Journal of Theoretical Biology* 254.4, pp. 843–849.

Horwich, Paul (2004). “A Use Theory of Meaning”. In: *Philosophy and Phenomenological Research* 68.2, pp. 351–372.

— (2008). “Wittgenstein’s Definition of ‘Meaning’ as ‘Use’”. In: *Annals of the Japan Association for Philosophy of Science* 16.1, pp. 133–141.

Hume, David (1738). *A Treatise of Human Nature.* Oxford: Clarendon Press.

Huttegger, Simon M. (2007a). “Evolution and the Explanation of Meaning”. In: *Philosophy of Science* 74.1, pp. 1–27.

— (2007b). “Evolutionary Explanations of Indicatives and Imperatives”. In: *Erkenntnis* 66.3, pp. 409–436.

— (2014). “How Much Rationality Do We Need to Explain Conventions?” In: *Philosophy Compass* 9.1, pp. 11–21.

Huttegger, Simon M., Brian Skyrms, Rory Smead, and Kevin J. S. Zollman (2009). "Evolutionary Dynamics of Lewis Signaling Games: Signaling Systems Vs. Partial Pooling". In: *Synthese* 172.1, pp. 177–191.

Huttegger, Simon M., Brian Skyrms, Pierre Tarrès, and Elliott Wagner (2014). "Some Dynamics of Signaling Games". In: *Proceedings of the National Academy of Sciences of the United States of America* 111.Suppl 3, pp. 10873–10880.

Huttegger, Simon M. and Kevin J. S. Zollman (2010). "Dynamic Stability and Basins of Attraction in the Sir Philip Sidney Game". In: *Proceedings of the Royal Society of London B: Biological Sciences* 277.1689, pp. 1915–1922.

— (2013). "Methodology in Biological Game Theory". In: *The British Journal for the Philosophy of Science* 64.3, pp. 637–658.

— (2016). "The Robustness of Hybrid Equilibria in Costly Signaling Games". In: *Dynamic Games and Applications* 6.3, pp. 347–358.

Jäger, Gerhard (2007). "The Evolution of Convex Categories". In: *Linguistics and Philosophy* 30.5, pp. 551–564.

Jäger, Gerhard, Lars P. Metzger, and Frank Riedel (2011). "Voronoi Languages: Equilibria in Cheap-talk Games With High-dimensional Types and Few Signals". In: *Games and Economic Behavior* 73.2, pp. 517–537.

Jäger, Gerhard and Robert van Rooij (2007). "Language Structure: Psychological and Social Constraints". In: *Synthese* 159.1, pp. 99–130.

Jiang, Junjie, Yu-Zhong Chen, Zi-Gang Huang, and Ying-Cheng Lai (2018). "Evolutionary Hypergame Dynamics". In: *Physical Review E* 98.4, p. 042305.

Jurafsky, Dan and James H. Martin (2009). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.* Second edition. Prentice Hall Series in Artificial Intelligence. Upper Saddle River, N.J: Prentice Hall.

Kahane, Guy, Edward Kanterian, and Oskari Kuusela, eds. (2007). *Wittgenstein and His Interpreters: Essays in Memory of Gordon Baker.* Reprint edition. Chichester: Wiley-Blackwell.

Kanazawa, Takafumi, Toshimitsu Ushio, and Tatsushi Yamasaki (2007). "Replicator Dynamics of Evolutionary Hypergames". In: *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans* 37.1, pp. 132–138.

Keefe, Rosanna and Peter Smith, eds. (1999). *Vagueness: A Reader.* Paperback edition. Cambridge, Massachusetts: MIT Press.

Klement, Kevin (2016). "Russell's Logical Atomism". In: *The Stanford Encyclopedia of Philosophy.* Ed. by Edward N. Zalta. Spring 2016. Metaphysics Research Lab, Stanford University.

Kohn, Marek (2008). “Darwin 200: The Needs of the Many”. In: *Nature News* 456.7220, pp. 296–299.

Kramer, Jos and Joël Meunier (2016). “Kin and Multilevel Selection in Social Evolution: A Never-ending Controversy?” In: *F1000Research* 5.

Kraugerud, Hanne A. and Bjørn T. Ramberg (2010). “The New Loud: Richard Rorty, Quietist?” In: *Common Knowledge* 16.1, pp. 48–65.

Kripke, Saul A. (1972). “Naming and Necessity”. In: *Semantics of Natural Language.* Ed. by Donald Davidson and Gilbert Harman. Synthese Library. Dordrecht: Springer Netherlands, pp. 253–355.

— (1982). *Wittgenstein on Rules and Private Language: An Elementary Exposition.* Cambridge: Harvard University Press.

Lassiter, Daniel and Noah D. Goodman (2017). “Adjectival Vagueness in a Bayesian Model of Interpretation”. In: *Synthese* 194.10, pp. 3801–3836.

Lawry, Jonathan and Oliver James (2017). “Vagueness and Aggregation in Multiple Sender Channels”. In: *Erkenntnis* 82.5, pp. 1123–1160.

Lewis, David (1969). *Convention: A Philosophical Study.* Cambridge: Harvard University Press.

Lipman, Barton L. (2009). “Why is Language Vague?” Unpublished.

Loreto, Vittorio, Andrea Baronchelli, and Andrea Puglisi (2010). “Mathematical Modeling of Language Games”. In: *Evolution of Communication and Language in Embodied Agents.* Ed. by Stefano Nolfi and Marco Mirolli. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 263–281.

Malachowski, Alan (2002). *Richard Rorty.* Philosophy Now. Chesham: Acumen.

Martínez, Manolo (2019). “Deception as Cooperation”. In: *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, p. 101184.

Martínez, Manolo and Peter Godfrey-Smith (2016). “Common Interest and Signaling Games: A Dynamic Analysis”. In: *Philosophy of Science* 83.3, pp. 371–392.

Maynard Smith, John (1982). *Evolution and the Theory of Games.* Cambridge: Cambridge University Press.

— (1991). “Honest Signalling: The Philip Sidney Game”. In: *Animal Behaviour* 42.6, pp. 1034–1035.

— (1994). “Must Reliable Signals Always Be Costly?” In: *Animal Behaviour* 47.5, pp. 1115–1120.

McDowell, John (2009). “Wittgensteinian “Quietism””. In: *Common Knowledge* 15.3, pp. 365–372.

McGinn, Marie (2011). “Grammar in the *Philosophical Investigations*”. In: *The Oxford Handbook of Wittgenstein*. Ed. by Oskari Kuusela and Marie McGinn. Oxford University Press, pp. 646–666.

Miller, Alexander and Crispin Wright, eds. (2002). *Rule-Following and Meaning*. Chesham: Acumen.

Morris, Katherine (2007). “Wittgenstein’s Method: Ridding People of Philosophical Prejudices”. In: *Wittgenstein and His Interpreters: Essays in Memory of Gordon Baker*. Ed. by Guy Kahane, Edward Kanterian, and Oskari Kuusela. Reprint edition. Chichester: Wiley-Blackwell, pp. 66–87.

Mühlenbernd, Roland (2011). “Learning With Neighbours”. In: *Synthese* 183.1, pp. 87–109.

— (2017). “The Change of Signaling Conventions in Social Networks”. In: *AI & SOCIETY*, pp. 1–14.

Mühlenbernd, Roland and Michael Franke (2012). “Signaling Conventions: Who Learns What Where and When in a Social Network?” In: *The Evolution of Language*. World Scientific, pp. 242–249.

Nersessian, Nancy J. (1992). “In the Theoretician’s Laboratory: Thought Experimenting as Mental Modeling”. In: *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association* 1992.2, pp. 291–301.

Nersessian, Nancy J. and Miles MacLeod (2017). “Models and Simulations”. In: *Springer Handbook of Model-Based Science*. Springer Handbooks. Cham: Springer, pp. 119–132.

Nowak, Martin A. and David C. Krakauer (1999). “The Evolution of Language”. In: *Proceedings of the National Academy of Sciences* 96.14, pp. 8028–8033.

Nowak, Martin A. and Karl Sigmund (2004). “Evolutionary Dynamics of Biological Games”. In: *Science* 303.5659, pp. 793–799.

O’Connor, Cailin (2014a). “Evolving Perceptual Categories”. In: *Philosophy of Science* 81.5, pp. 840–851.

— (2014b). “The Evolution of Vagueness”. In: *Erkenntnis* 79.4, pp. 707–727.

— (2015). “Ambiguity is Kinda Good Sometimes”. In: *Philosophy of Science* 82.1, pp. 110–121.

— (2017). “Evolving to Generalize: Trading Precision for Speed”. In: *The British Journal for the Philosophy of Science* 68.2, pp. 389–410.

— (2019). “Games and Kinds”. In: *The British Journal for the Philosophy of Science* 70.3, pp. 719–745.

Okasha, Samir (2006). *Evolution and the Levels of Selection*. Oxford University Press.

Pacheco, Jorge M., Vítor V. Vasconcelos, Francisco C. Santos, and Brian Skyrms (2015). “Co-evolutionary Dynamics of Collective Action with Signaling for a Quorum”. In: *PLOS Computational Biology* 11.2. Ed. by Carl T. Bergstrom, e1004101.

Pagel, Mark (2009). “Human Language as a Culturally Transmitted Replicator”. In: *Nature Reviews Genetics* 10.6, pp. 405–415.

Parisi, Domenico (2004). “Language as Pragmatics: Studying Meaning With Simulated Language Games”. In: *Seduction, Community, Speech: A Festschrift for Herman Parret*. Ed. by Frank Brisard, Michael Meeuwis, and Bart Vandenabeele, pp. 139–149.

Pichler, Alois (2007). “The Interpretation of the *Philosophical Investigations*: Style, Therapy, *Nachlass*”. In: *Wittgenstein and His Interpreters: Essays in Memory of Gordon Baker*. Ed. by Guy Kahane, Edward Kanterian, and Oskari Kuusela. Reprint edition. Chichester: Wiley-Blackwell, pp. 123–144.

Plant, Bob (2004). “The End(s) of Philosophy: Rhetoric, Therapy and Wittgenstein’s Pyrrhonism”. In: *Philosophical Investigations* 27.3, pp. 222–257.

Portner, Paul (2005). *What is Meaning? Fundamentals of Formal Semantics*. Fundamentals of Linguistics. Malden: Blackwell Publishers.

Proops, Ian (2017). “Wittgenstein’s Logical Atomism”. In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Winter 2017. Metaphysics Research Lab, Stanford University.

Putnam, Hilary (1970). “Is Semantics Possible?” In: *Metaphilosophy* 1.3, pp. 187–201.

— (1975). “The Meaning of ‘Meaning’”. In: *Mind, Language and Reality*. Vol. 2. Philosophical Papers. Cambridge: Cambridge University Press.

— (1994). *Pragmatism: An Open Question*. John Wiley & Sons, Inc.

Qing, Ciyang and Michael Franke (2014). “Gradable Adjectives, Vagueness, and Optimal Language Use: A Speaker-oriented Model”. In: *Proceedings of the 24th Semantics and Linguistic Theory Conference (SALT 24)*. Ed. by Todd Snider, Sarah D’Antonio, and Mia Weigand. LSA and CLC Publications, pp. 23–41.

Quine, Willard Van Orman (1936). “Truth by Convention”. In: *Journal of Symbolic Logic*, pp. 77–106.

Ratner, Nancy and Jerome Bruner (1978). “Games, Social Exchange and the Acquisition of Language”. In: *Journal of Child Language* 5.3, pp. 391–401.

Rescher, Nicholas (2005). *What If? Thought Experimentation in Philosophy*. New Brunswick: Transaction Publishers.

Rorty, Richard, ed. (1967). *The Linguistic Turn: Recent Essays in Philosophical Method.* Chicago; London: The University of Chicago Press.

— (1976). “Keeping Philosophy Pure”. In: *The Yale Review* LXV.3, pp. 336–356.

— (1979). *Philosophy and the Mirror of Nature.* Oxford: Basil Blackwell.

— (1982a). *Consequences of Pragmatism: Essays, 1972-80.* Minneapolis: University of Minnesota Press.

— (1982b). “Keeping Philosophy Pure: An Essay on Wittgenstein”. In: *Consequences of Pragmatism: Essays, 1972-80.* Minneapolis: University of Minnesota Press, pp. 19–36.

— (1989). *Contingency, Irony, and Solidarity.* Cambridge: Cambridge University Press.

— (1991). “Wittgenstein, Heidegger, and the Reification of Language”. In: *Essays on Heidegger and Others.* Vol. 2. Philosophical Papers. Cambridge University Press.

— (2007). “Wittgenstein and the Linguistic Turn”. In: *Philosophy as Cultural Politics.* Philosophical Papers 4. Cambridge: Cambridge University Press, pp. 160–175.

Roth, Alvin E. and Ido Erev (1995). “Learning in Extensive-form Games: Experimental Data and Simple Dynamic Models in the Intermediate Term”. In: *Games and Economic Behavior* 8.1, pp. 164–212.

Russell, Bertrand (1905). “On Denoting”. In: *Mind* 14.56, pp. 479–493.

— (1919). “The Philosophy of Logical Atomism”. In: *The Monist* 29.3, pp. 345–380.

— (1923). “Vagueness”. In: *Australasian Journal of Psychology and Philosophy* 1.2, pp. 84–92.

Ryle, Gilbert (1962). “Abstractions”. In: *Dialogue: Canadian Philosophical Review/Revue Canadienne de Philosophie* 1.1, pp. 5–16.

Sanborn, Adam N. and Nick Chater (2016). “Bayesian Brains without Probabilities”. In: *Trends in Cognitive Sciences* 20.12, pp. 883–893.

Santana, Carlos (2014). “Ambiguity in Cooperative Signaling”. In: *Philosophy of Science* 81.3, pp. 398–422.

Sasaki, Yasuo and Kyoichi Kijima (2012). “Hypergames and Bayesian Games: A Theoretical Comparison of the Models of Games With Incomplete Information”. In: *Journal of Systems Science and Complexity* 25.4, pp. 720–735.

— (2016). “Hierarchical Hypergames and Bayesian Games: A Generalization of the Theoretical Comparison of Hypergames and Bayesian Games Considering Hierarchy of Perceptions”. In: *Journal of Systems Science and Complexity* 29.1, pp. 187–201.

Schlag, Karl H. (1998). “Why Imitate, and If So, How? A Boundedly Rational Approach to Multi-armed Bandits”. In: *Journal of Economic Theory* 78.1, pp. 130–156.

Sillari, Giacomo (2013). “Rule-following as Coordination: A Game-theoretic Approach”. In: *Synthese* 190.5, pp. 871–890.

Simon, Herbert A. (1955). “A Behavioral Model of Rational Choice”. In: *The Quarterly Journal of Economics* 69.1, pp. 99–118.

— (1986). “Rationality in Psychology and Economics”. In: *The Journal of Business* 59.4, S209–S224.

Skyrms, Brian (1996). *Evolution of the Social Contract.* Cambridge: Cambridge University Press.

— (2009). “Evolution of Signalling Systems With Multiple Senders and Receivers”. In: *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 364.1518, pp. 771–779.

— (2010). *Signals: Evolution, Learning, and Information.* Oxford: Oxford University Press.

Smaldino, Paul E. (2017). “Models Are Stupid, and We Need More of Them”. In: *Computational Social Psychology.* Ed. by Robin R. Vallacher, Stephen J. Read, and Andrzej Nowak. New York: Routledge, pp. 311–331.

Sorensen, Roy (2009). “sorites arguments”. In: *A Companion to Metaphysics.* Ed. by Jaegwon Kim, Ernest Sosa, and Gary S. Rosenkrantz. Second edition. John Wiley & Sons, Inc., pp. 565–566.

Speaks, Jeff (2018). “Theories of Meaning”. In: *The Stanford Encyclopedia of Philosophy.* Ed. by Edward N. Zalta. Winter 2018. Metaphysics Research Lab, Stanford University.

Spence, Michael (1978). “Job Market Signaling”. In: *Uncertainty in Economics.* Ed. by Peter Diamond and Michael Rothschild. Academic Press, pp. 281–306.

— (2002). “Signaling in Retrospect and the Informational Structure of Markets”. In: *American Economic Review* 92.3, pp. 434–459.

Spike, Matthew, Kevin Stadler, Simon Kirby, and Kenny Smith (2017). “Minimal Requirements for the Emergence of Learned Signaling”. In: *Cognitive Science* 41.3, pp. 623–658.

Steels, Luc (1995). “A Self-Organizing Spatial Vocabulary”. In: *Artificial Life* 2.3, pp. 319–332.

Stern, David G. (2004). *Wittgenstein’s Philosophical Investigations: An Introduction.* New York: Cambridge University Press.

— (2007). “The Uses of Wittgenstein’s Beetle: *Philosophical Investigations* §293 and Its Interpreters”. In: *Wittgenstein and His Interpreters: Essays in Memory of Gordon Baker*. Ed. by Guy Kahane, Edward Kanterian, and Oskari Kuusela. Reprint edition. Chichester: Wiley-Blackwell, pp. 248–268.

Stokhof, Martin (2013). “Formal Semantics and Wittgenstein: An Alternative?” In: *The Monist* 96.2, pp. 205–231.

Taylor, Peter D. and Leo B. Jonker (1978). “Evolutionary Stable Strategies and Game Dynamics”. In: *Mathematical Biosciences* 40.1, pp. 145–156.

Thompson, Bill, Simon Kirby, and Kenny Smith (2016). “Culture Shapes the Evolution of Cognition”. In: *Proceedings of the National Academy of Sciences* 113.16, pp. 4530–4535.

Tomasello, Michael (2008). *Origins of Human Communication*. The Jean Nicod Lectures 2008. Cambridge: MIT Press.

Train, Kenneth E. (2009). *Discrete Choice Methods with Simulation*. Cambridge: Cambridge University Press.

Travis, Charles (2006). *Thought’s Footing: A Theme in Wittgenstein’s Philosophical Investigations*. New York: Oxford University Press.

Unger, Peter (1979). “There Are No Ordinary Things”. In: *Synthese* 41.2, pp. 117–154.

Vanderschraaf, Peter (1998). “The Informal Game Theory in Hume’s Account of Convention”. In: *Economics and Philosophy* 14.02, pp. 215–247.

Vehtari, Aki and Janne Ojanen (2012). “A Survey of Bayesian Predictive Methods for Model Assessment, Selection and Comparison”. In: *Statistics Surveys* 6, pp. 142–228.

Vul, Edward, Noah Goodman, Thomas L. Griffiths, and Joshua B. Tenenbaum (2014). “One and Done? Optimal Decisions From Very Few Samples”. In: *Cognitive Science* 38.4, pp. 599–637.

Wagner, Elliott (2009). “Communication and Structured Correlation”. In: *Erkenntnis* 71.3, pp. 377–393.

— (2012). “Deterministic Chaos and the Evolution of Meaning”. In: *The British Journal for the Philosophy of Science* 63.3, pp. 547–575.

— (2013). “The Dynamics of Costly Signaling”. In: *Games* 4.2, pp. 163–181.

Waller, Bruce (1977). “Chomsky, Wittgenstein, and the Behaviorist Perspective on Language”. In: *Behaviorism* 5.1, pp. 43–59.

Watts, Duncan J. and Steven H. Strogatz (1998). “Collective Dynamics of ‘Small-world’ Networks”. In: *Nature* 393.6684, pp. 440–442.

Wickham, Hadley (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wilensky, Uri and William Rand (2015). *An Introduction to Agent-Based Modeling: Modeling Natural, Social, and Engineered Complex Systems with NetLogo*. Cambridge: The MIT Press.

Wilson, David Sloan (1975). “A Theory of Group Selection”. In: *Proceedings of the National Academy of Sciences of the United States of America* 72.1, pp. 143–146.

Winter, Yoad (2016). *Elements of Formal Semantics: An Introduction to the Mathematical Theory of Meaning in Natural Language*. Edinburgh Advanced Textbooks in Linguistics. Edinburgh: Edinburgh University Press.

Wisdom, John (1953). *Philosophy and Psycho-analysis*. Basil Blackwell.

Wittgenstein, Ludwig (1922). *Tractatus Logico-Philosophicus*. Trans. by C. K. Ogden. London: Kegan Paul, Trench and Trübner.

— (1953). *Philosophical Investigations*. Oxford: Basil Blackwell.

— (1958). *The Blue and Brown Books: Preliminary Studies for the ‘Philosophical Investigations’*. Oxford: Basil Blackwell.

— (2002). *The Blue and Brown Books: Preliminary Studies for the ‘Philosophical Investigations’*. Second edition, reprinted. Oxford: Basil Blackwell.

— (2009). *Philosophical Investigations*. Ed. by P. M. S. Hacker and Joachim Schulte. Trans. by G. E. M. Anscombe, P. M. S. Hacker, and Joachim Schulte. Revised 4th edition. Chichester: Wiley-Blackwell.

Wright, Crispin (2007). “Rule-Following without Reasons: Wittgenstein’s Quietism and the Constitutive Question”. In: *Ratio* 20.4, pp. 481–502.

Zahavi, Amotz (1975). “Mate Selection—A Selection for a Handicap”. In: *Journal of Theoretical Biology* 53.1, pp. 205–214.

— (1977). “The Cost of Honesty: Further Remarks on the Handicap Principle”. In: *Journal of Theoretical Biology* 67.3, pp. 603–605.

Zollman, Kevin J. S. (2005). “Talking to Neighbors: The Evolution of Regional Meaning”. In: *Philosophy of Science* 72.1, pp. 69–85.

— (2011). “Separating Directives and Assertions Using Simple Signaling Games”. In: *The Journal of Philosophy* 108.3, pp. 158–169.

Zollman, Kevin J. S., Carl T. Bergstrom, and Simon M. Huttegger (2013). “Between Cheap and Costly Signals: The Evolution of Partially Honest Communication”. In: *Proceedings of the Royal Society B: Biological Sciences* 280.1750, p. 20121878.

Zollman, Kevin J. S. and Rory Smead (2010). “Plasticity and Language: An Example of the Baldwin Effect?” In: *Philosophical Studies* 147.1, pp. 7–21.