Journal of Language Evolution, 2021, 136–153
doi: 10.1093/jole/lzab003
Advance Access Publication Date: 3 November 2021
Research article

The philosophical interpretation of language game theory

Nick Zangwill [1,2]

[1]Department of Philosophy, University College London, UK and [2]Department of Philosophy, University of Lincoln, Lincoln, UK
*Corresponding author: zangwillnick@gmail.com

Abstract

I give an informal presentation of the evolutionary game theoretic approach to the conventions that constitute linguistic meaning. The aim is to give a philosophical interpretation of the project, which accounts for the role of game theoretic mathematics in explaining linguistic phenomena. I articulate the main virtue of this sort of account, which is its psychological economy, and I point to the casual mechanisms that are the ground of the application of evolutionary game theory to linguistic phenomena. Lastly, I consider the objection that the account cannot explain predication, logic, and compositionality.

Key words: language games; game theory; evolution; interdisciplinarity; meaning

1. Introduction: interpreting language game theory

Evolutionary game theory has made impressive strides in modeling many aspects of natural language. Of particular interest is the light it has cast on the origin and nature of the conventions that constitute basic word meanings—for example, that the English word 'dog' refers to dogs, or that 'blue' refers to blueness. Evolutionary game theory has been used to model the cultural evolutionary processes that generate conventions in order to give an explanation both of how basic linguistic meaning arises, and also an account of what it is. (For some examples of an extensive genre, see Nowak and Krakauer 1999; Nowak, Plotkin and Krakauer 1999; Hofbauer and Huttegger 2008; and Argiento et al. 2009).

However, in my view, the complexity of the application of the theory and the abstractness of its mathematical formulations has meant that its distinctive philosophical contribution has been under-appreciated. It has made little difference to most of those working in the philosophy of language.[1] In this article, I aim to articulate the main philosophical benefits of the program. In particular, the program is of philosophical interest primarily because it gives an explanation, in mathematical detail, of how a word comes to refer to a particular thing or feature of the world. Furthermore, as we shall see, the explanation is relatively sparse in what it posits in the psychology of language users. The approach also has an empirical predictive aspect to it, such as its use in modeling language drift; and it has been put to use to explain other linguistic phenomena, such as aspects of pragmatics, but I pass over these endeavors, which depend on basic word meaning—reference relations—being in place. So, the focus will be primarily on the explanation of basic referential relationships, such as that 'dog' refers to dogs and 'blue' refers to blueness.

Although we might be impressed by the formal power of evolutionary game theory, we need to ensure that the somewhat abstract mathematical formulae are grounded in empirically tractable reality. After all, the mathematical principles of evolutionary game theory are being deployed in order to understand real-world phenomena. If so, we need

to know *how* the mathematics has application to those phenomena. That means that the mathematical formal descriptions need what we can call an 'interpretation'. In 1991, Ariel Rubinstein made this point forcefully about game theory in general, when he wrote:

I believe that discussion or application of game theory is utterly meaningless without a proper interpretation.

The reason for this is that

. . . game theory is not simply a matter of abstract mathematics but concerns the real world.

As game theory is at once abstract and concrete, we must build a bridge between the abstract formal concepts of the theory and reality (All quotations Rubinstein 1991: 909).

Rubinstein's forthright and insightful complaints apply in spades to work on the application of evolutionary game theory to language. I aim to provide at least the outlines of an answer to a Rubinsteineque challenge in this area. We need to build Rubinstein's bridge.

The general point here is that one cannot just describe a mathematical structure and then just *declare* that it is explanatory, not even if there is a good mapping between the mathematical structure and concrete phenomena. Only if mathematical structures have a plausible interpretation, such that they can be seen to correspond to real structures of the phenomena in question, do we have solid explanatory progress. Only then can the mathematical structures be part of the explanation of a concrete phenomenon in the way that many people think that mathematical explanations are often part of good explanations of nonmathematical phenomena.[2] Then we would know *why* the mapping obtains.

Thus, the concern will be with ensuring that the evolutionary game theoretic approach to language has a proper grounding. The point of that, from a philosophical point of view, is that once it is in place, can we then articulate clearly the philosophical benefits of the approach; and we can then also see what remains to be done by way of filling out the program, and extending it to various other phenomena of language. The formalisms need philosophy and philosophy needs the formalisms. However, to date, while there has been a rich exploration of the formalisms (e.g. Huttegger 2007), there has been less than is needed in the way of philosophical interpretation.

Interpretation is needed because some philosophers will rightly view the technical results with suspicion until they can be underwritten by a convincing philosophical interpretation, which speaks to what in the world enacts the game theoretical structures. This is why merely recapitulating, for philosophers, some of the technical and predictive achievements previously obtained by others in other disciplines is not the best way to sell the evolutionary game theoretic approach to language to philosophers. Rubinstein's bridge remains to be built.

In the background of the discussion to follow lies a general conception of interdisciplinarity to which I incline, whereby those in one discipline do not merely strive to replicate what is done in another discipline. Instead, there should be a respect for distinct disciplinarily approaches to the same or similar subject matters. That way something new can be built with the two distinct disciplines working together, rather than each one trying to occupy the space of the other. For this reason, philosophers are probably better off not pursuing the mathematics of evolutionary game theory, which others are better trained to do; instead, they can aspire to illuminate what grounds the enterprise. Only then can we appraise the impact of the program on traditional and current issues in the philosophy of language. In other words, we seek an interpretation, in Rubinstein's sense, as a way of assessing and extending its impact. (I return to these thoughts about interdisciplinarity in Section 3, and in section 9 at the end of this article.)

I proceed by situating the understanding that the evolutionary language game theory approach to language yields in the context of a number of standard philosophical issues about language; and I shall underline what I take to be the main significance or benefit of the approach, which is its explanatory economy. In this respect, there is, as we shall see, a significant contrast with other accounts that invoke complex mental states that are about mental states (sometimes called 'metarepresentations'). I outline the tangible structures that ground the mathematical explanations, which, as we will see, turn out to be in part nested arrays of psychological states of a certain kind that stand in dispositional causal relations to each other. Lastly, I turn to some central aspects of language that have received little attention within the program, and I show that they are not in principle intractable. In particular, issues of semantic and logical structure might be seen as problematic for the approach, even as an objection to the entire approach, given the systematic nature of language. I show that these issues can in principle be addressed within the evolutionary game theoretic framework, and I make some positive suggestions about this matter. I end with some comments about the whole enterprise.

2. Game theory and language evolution

Let us begin by going more slowly than is usual over the basic building blocks of the application of evolutionary language game theory to language. A primal question is: how does language (public language[3]) come into being? In considering this question, we put aside the question of how we can *think* about anything. That is taken as given. Settling that question leaves open many important questions about language (see Lowe 1996, chapter 5; contrast McGinn 1984: 144–151). *Assuming* that we have thoughts about things, how is linguistic meaning generated? In particular, how is the connection set up between words, as perceivable physical symbols, on the one hand, and objects and properties, on the other hand? Given that we can *think* about dogs, how does it come about that a word, such as the English word 'dog', refers to dogs? That is nontrivial. On the evolutionary game theoretic approach, the process by which 'dog' comes to refer to dogs is not irrelevant history but part of what it is for 'dog' to refer to dogs—even though the process is not transparent to those who engage in linguistic behavior. The process by which linguistic meaning arises is part of what it is. Along with other artifacts, the nature of words, and what words mean, is given in part by their history. This means that we should ask: what is the historical process by which words come to mean what they do?

The basic idea of the evolutionary game theoretic approach to language is that there are 'players' who send and receive signals to and from each other, with benefits and costs (or probabilities of benefits and costs) that are consequential on whether or not senders and receivers both associate the signal with the same thing or property. The game theoretic account assigns payoffs, positive, and negative, for coinciding and failing to coincide on the meanings (references to objects or properties) of linguistic symbols. (See Lewis 1969; and more recently Millikan 2002, and Skyrms 2010). Within a coordinating group, when people produce a word (a perceivable physical symbol) and associate it with a thing or property, and other people—those who consume what is produced—associate it with the same thing or property, then there is a positive payoff, and if not, not. These payoffs are represented in matrixes, which may be mathematically manipulated in order to generate complex behavior of a collective of players from facts concerning individual players. Now, in a speech community, the members of the group are assumed to have symmetrical payoffs; that is, they have coinciding interests—what benefits one benefits the other, and what harms one harms the other. Degrees of positive and negative payoffs may be added to this framework. There may be proportionality between degrees of agreement and the amount of payoff, so that a word-world mapping reproduces itself in proportion to the degree of agreement. If players assign the same meaning to the symbols, or more or less the same meaning, then, in one sense, we may say that 'information' is transmitted; and that is advantageous—and usually mutually advantageous in such a group—given the causal role of the thing or property that is thought of, and which is associated with the symbol. Given the payoffs, such assignments of meanings to symbols are more likely to survive into the next round of the signaling game, and a divergence in assignment is less likely to survive. There are circumstances where payoffs are not symmetrical. In such cases, deception strategies might emerge. But they can only do so against a background of agreed meanings within a community where there are mostly symmetrical payoffs.

Typically, multiple players engage in the signaling by which linguistic conventions are generated. Convergence in linguistic behavior arises, and a linguistic convention is established, where there are stable solutions to coordination problems where multiple players engage in signaling behavior with other players in a dynamically changing situation. This typically delivers an 'evolutionarily stable strategy', in John Maynard-Smith's terms, where an 'evolutionarily stable strategy' is one such that nearby strategies will converge, or would converge, on it, because it can survive competition with a range of alternative mutant strategies (Maynard-Smith's 1982). Applied to linguistic meaning, the idea, very roughly, is that when and only when players converge on such an evolutionarily stable strategy, which yields a stable syndrome of behaviors, then the word—that is, the perceivable physical symbol—refers to what the players think of when they deploy the symbol.

Of course, not everything that players think of is part of the meaning of the terms in question. When hearing the name 'John', one might think of John's nose or John's brother. Nevertheless, certain thoughts are canonical, in that they are functionally appropriate given the convention: thoughts about John, for example. Just as the linguistic conventions themselves are cultural artifacts, so some psychological responses during enactments of the convention are more or less fitting to the cultural artifact, just as a bicycle is an artifact and there is bicycle-appropriate behavior that fits the bicycle-artifact. Such mutual fitting is what a linguistic convention *is*. The symbol refers to what it does only given situations of dynamically stable convergence among players. And the evolutionary game theoretical approach describes *how* this convergence is achieved (which certainly helps with understanding how it is possible!). It is

how the word 'dog' comes to refer to dogs, and it is what it *is* for 'dog' to refer to dogs. Furthermore, there is no reason why these symbol-world reference relations might not be more or less determinate: one could have symbols for poodle, dog, and animal.

This signaling approach to linguistic meaning is an evolutionary account in a broad sense of the word 'evolutionary'. An 'evolutionary system' in this broader sense is not necessarily biological, but it must include replicators and selection mechanisms, which means that there is a feedback structure such that there is an increase in growth rate consequential on successful strategies. That is, these strategies reproduce themselves to a degree corresponding to their success (ignoring extraneous factors). So, the population of replicators in a generation depends on the success of replication strategies in the previous round of the game. In a *biological* evolutionary context, where a game is often a life, what replicates are phenotypes, and fitness is a property of phenotypes. In linguistic cultural evolution, what replicates are mappings of objects or properties to physical symbols. And evolutionary game theory describes (and predicts) the way convergence in mappings is achieved and is likely to be sustained in groups of players under certain conditions.

The convergence on meanings has a describable mathematical structure. What is crucial is that what generates meanings for symbols in dynamical systems are stable rest points that emerge and that are accessible from nonrest points. (Some but not all of these rest points are maximally efficient 'Nash equilibria' (Nash 1950)[4]). A stable strategy, or conjunction of stable strategies, on which there is a tendency to converge, can be mathematically representable as an 'attractor'. Such an attractor may not be the most efficient signaling convention possible (Pawlowitsch 2008). But it needs to be one that is stable within certain parameters, and one that is *accessible* from previous conditions.

Repeated signaling games generate complex structures given the various possible combinations of actions of the players, and outcomes that depend on the actions of other players, and of outcomes of actions that depend on the outcomes of previous interactions. These complexities are representable in a dynamic extension of classical game theory (*locus classicus* Morgenstern and Von Neumann 1944). Applied to signaling, the game theoretic matrixes may be used to describe how linguistic conventions arise as a stable coordinated behavior pattern that is a relatively satisfactory solution to signaling problems. These solutions are arrived at given repeated rounds of the signaling game because solution strategies are encouraged by feedback mechanisms, whereas other strategies are discouraged. This process of dynamic convergence on a stable solution is mathematically describable and predictable (at least probabilistically) based on individual matrixes—assuming various idealizations.

These, very roughly, are the basic elements of the signaling evolutionary game theoretic account of language—'very roughly' because the mathematical technicalities have been omitted and also because, as is usual in science, the explanation is idealized in various respects (for example, no extraneous comets land from outer space, and there are no competing evolutionary dynamic pressures with quite different kinds of payoffs, which are in competition with the signaling game payoffs). I have given an informal description of the basic ideas of evolutionary language game theory, which has received a series of elegant formal expositions in the last generation. The mathematical results are readily available, and there is no point in recapitulating them here. What is left untouched by both informal and formal treatments are: (1) what the explanatory benefits are and (2) how to interpret the mathematical formalisms.[5]

3. Game theoretic explanation

Let us now turn to address the interpretation of the application of evolutionary language game theory to language. This application is a case of use of mathematics in the explanation of nonmathematical phenomena. Mark Colyvan has persuasively argued for the pervasiveness of this scenario (Colyvan 2001, chapter 3). There are many mathematical explanations of nonmathematical phenomena, such as biological phenomena of host–parasite ratios. Of course, where nonmathematical phenomena are explained by means of mathematics, it is never the *whole* explanation. Mathematics cannot explain concrete phenomena on its own; that would be extreme Pythagoreanism. But it is an essential part of the whole explanation.

The question is *how* mathematic facts, together with nonmathematical facts, can explain something. It seems that the application of mathematics to a causal system, in biology, linguistics, or human society, needs to avail itself of *some mechanism* by which the mathematics does its work (Machamer, Lindlay and Card 2000). In the case of evolutionary language game theory, in particular, this mechanism includes at least replicators and causal feedback loops; so we need to identify these. We need not conceive of this mechanism in a very strict way, so that it implies discrete contiguous parts of a thing that work together to produce an effect, as in many human-made machines (see Dupre 2017; Woodward 2013). Nevertheless, there must be some systematic causal basis of the feedback by which selection

of replicators occurs. The systematic causal basis, I suggest, must itself have dynamically stable properties.[6]

The thought that we need to identify a mechanism underpinning mathematical explanations of nonmathematical phenomena is connected with the kind of interdisciplinarity that I believe we need when thinking about evolutionary language game theoretic explanations, as was mentioned in the introduction, and as will be revisited in the coda. Doing more mathematics cannot solve the mechanism problem, and therefore does not engage with the fundamental philosophical issues raised by the application of evolutionary game theory to language. Without mechanism, mathematics is a wheel that spins idly. If mathematics is to function explanatorily with respect to concrete phenomena, there must be traction, as it were, such that the mathematics drives along physical processes, or at least directs them along certain pathways, rather than others. Traction means mechanism. And identifying mechanisms is not a job for mathematicians, but falls naturally within the domain of philosophy.

Now, evolution—whether biological or cultural—requires replication and selection, and these must be effective somehow if the game theoretic matrixes are not to remain a description of Platonic heaven rather than biological or cultural reality. At least, we can say that if the matrixes of game theory model the realities of biological or cultural evolution, then, corresponding to the matrixes, there are complex *conditional causal* facts about the players. There must be a mapping from conditional causal facts and causal structures onto the matrixes.

Consider some game theoretic structure, say, a familiar prisoner's dilemma matrix.

	y defects (D)	y cooperates (C)
x defects (D)	x gets 2/y gets 2	x gets 4/y gets 1
x cooperates (C)	x gets 1/y gets 4	x gets 3/y gets 3

Suppose that this matrix describes some biological or cultural phenomenon. A limitation, of course, is that this matrix is static, without temporal variables. These could be added; but doing so would add complexity. So, let us stay with a one-off game for the time being. Then, where A and B are types of actions, we have a reward ('R') table corresponding to the matrix.

If Cx and Dy, then R1x and R4y
If Dx and Dy, then R2x and R2y
If Dx and Cy, then R4x and R1y
If Cx and Cy, then R3x and R3y

These conditionals, however, are not fundamental; they hold in virtue of dispositional causal relations, and all of them together constitute a 'functional system', in one sense of 'functional'—the one popular in the philosophy of mind in the 1970s and 1980s, whereby it denotes a cluster of interlocking dispositional causal properties (Shoemaker 1981), and it has no historical implications.

Reward tables are generated by what are termed the *strategies* of the individual players. A strategy of player x might be, for example, the conjunction:

If Dy, then do Dx
If Cy, then do Dx

That is stated as a pair of prescriptions, which one might self-consciously follow—actions conditional on other's actions. Or we can take it as a statement of a conditional fact, if Dy then Dx (other things being equal), and if Cy, then Dx (other things being equal). That strategy, when pursued by multiple players generates prisoner's dilemma matrixes. Let us not worry yet about exactly what strategies are; they are, at least, conditional prescriptions concerning actions or facts about how a player would act under certain conditions.

If we now add a dynamic aspect, where games are repeated, then rewards R1–R4 generate differential rates of strategy reproduction, biological or cultural, which feed into the next rounds of the game. *What* is replicated, it seems, are strategies, or rather types of strategies, where strategies look like properties of objects (players or collections of players). However, evolution—biological or cultural—can only happen if strategies can vary between rounds of the game. This means that strategies are not best thought of as properties of objects that persist from one round to another. As Richard Dawkins emphasized in the biological domain (Dawkins 1976), what persists from generation to generation are not organism tokens. Strategies are modified according to the differential rewards that they incur. Strategies themselves, in that sense, may persist in similar or altered form down the generations. So, it may be better to think of strategies themselves as what persists.[7]

Repeating games allow for the evolution of strategies, as replication depends on rewards determining rates of propagation from one generation to the next. So, for example, Robert Axelrod's famous 'Tit for Tat' strategy, or one of its descendants, might emerge after a significant number of rounds of computer simulated evolutionary design (Axelrod 1984; see also Dawkins 1986, chapter 2).

Now, an individual player's strategy contributes to a causal structure, one that, in effect, computes outcomes given multiple players. The conditionals of reward tables hold in virtue of causal dispositions generated by all the individual strategies of the interacting parties, acting in concert. The overall causal structure is determined by the strategies of the players who mutually interact, and that ultimately explains why the matrixes hold. That is, the matrixes of game theory, when used to explain concrete phenomena, describe nested causal conditionals that are generated by the strategies of a set of players, where the relevant set of players are those whose outcomes depend on the strategies of other players in the set. (Causally isolated players can be ignored.[8]) The injection of dynamic variables corresponds to the mutation of casual structures over time as those strategies change under selective pressure. This is the metaphysical ground of the matrixes, where the 'metaphysical ground' means the reality in virtue of which they hold.

Nevertheless, the pure mathematics that describes the overall causal system does explanatory work; it contributes ineliminably to the explanation. There can be common mathematical explanations of phenomena that are realized in radically different causal structures—in biological evolution that produces organisms, and in cultural evolution that produces economic or linguistic behavior, for example. That is why we need the mathematical level of explanation, one which cuts across differences in realization. There are salient explanatory uniformities that are not captured at the level of realization, and that are only captured by the mathematical description. There are explanatory losses without the mathematical description.[9]

4. Explanation and psychological austerity

Let us now turn to consider a major virtue of the evolutionary game theoretic account of language. This virtue is its explanatory economy. In particular, part of the elegance of the evolutionary game theoretic account of language games is that linguistic meaning need not derive from complex (higher-order) mental states about mental states. Paul Grice required that speakers intend to cause beliefs or other responses in an audience who have beliefs about those intentions of the speaker (Grice 1989; see also Schiffer 1972; Sperber and Wilson 1986). Not all of Grice's views have been influential, but the idea that language use depends on beliefs and intentions about other minds that recognize those beliefs and intentions has been so influential that it sometimes is even taken to be a platitude in little need of justification. Nevertheless, it is mistaken. For, evolutionary language game theory shows that there can be convergence on linguistic conventions without the parties converging because players intend or believe that other players intend or believe something about their intentions or beliefs. A player needs to represent other players as behaving in various ways, and perhaps as referring to objects and properties, but not as thinking about the first players' intentions and beliefs. Furthermore, and I expand on this more in the next section, evolutionary game theory explains how Gricean sophistication, where it exists, is possible.[10]

The evolutionary game theoretic approach to language is also an alternative to views that assume that speakers and their audience have specifically linguistic or semantic knowledge of principles like ['p' means S] (see, e.g. Wiggins 1997). Players can converge in word-world mappings without having knowledge or even beliefs about what others mean by a word or about the meanings of words. Alternatively, a theory might require knowledge of axioms of truth theory, such as ['Snow is white' is true if and only if snow is white] (Davidson 1982). One semantic view requires that speakers and hearers deploy a notion of meaning while the other requires that they deploy a notion of truth. In contrast, evolutionary game theory allows that people can master and deploy language without beliefs with either kind of semantic content. To converge on linguistic conventions that establish reference relations, one need not have beliefs about other beliefs, or semantic beliefs (about meaning or truth), or beliefs about the speaker's semantic intentions. Nevertheless, it *is* plausible that speakers do deploy a basic notion of *reference* in their intentions. But that falls far short of the Gricean psychological arsenal attributed to language users.

We might put the general negative point, echoing Ludwig Wittgenstein, by saying that practicing as we do when we follow a linguistic convention is just our 'form of life'. One idea here is that the players do not have *reasons* for using language as they do. Sometimes we may have reasons and sometimes we may have sophisticated intentions and beliefs about other minds—but often we do not. This lack of reasons was central to Wittgenstein's *sprach-spiel* account of language, which he developed in the early 1930s (e.g. in Wittgenstein 1953, 1958), which is the precursor to the later technical development of evolutionary game theoretic approaches to language in the 1980s and 1990s. The word 'spiel' in German does not quite translate into 'game' in English, as it also has the sense of 'skit' or 'play', as in a theater production, or when playing the game of charades. The German word 'spiel' captures more of the acting or playing or play-acting aspect of the social interactions that produce convergence on the ritualistic or semi-institutionalized behavior that constitutes linguistic meaning. (Perhaps 'language-

play' would be an alternative English translation than 'language game'). The mathematical game theoretic approach is an extension of Wittgenstein's *sprach-spiel* approach.[11] There is a mechanism encouraging convergence without players converging because they have reasons to do so. They are trained or drilled; it is a nonrational process.

The lack of reasons is connected with the kind of knowledge that is in play. On the game theoretic account, the use of language is a practical ability that need not be, and rarely is, articulable in terms of the possession of propositional knowledge or beliefs about other minds or beliefs about meanings or truth. Compare our understanding of meaningful gestures, such as the difference between an ordinary handshake and a 'high-five'. There is a difference between these two meaningful actions that most of us know, but it would not be easy to say exactly what it is. Nonetheless, we have *practical knowledge* of the difference between a handshake and a high-five without *propositional knowledge* of the meaning of each of these acts, or of the mental life of other parties to these gestures. In contrast, the Gricean would require a lot of sophisticated propositional knowledge to do a handshake or a high-five, which we surely lack. Language can proceed with practical knowledge, rather than sophisticated propositional knowledge. In this respect linguistic activity is like other ritualistic behavior (Wittgenstein 1993).

Apart from Wittgenstein, for a long time, only David Lewis explored the game theoretic approach to language. Although broadly on the right lines in invoking Thomas Schelling's work in game theory, Lewis failed to avoid the Gricean mistake of attributing too much psychological sophistication to players in his account of the 'common knowledge' that language users have. Like Grice, Lewis attributed implausibly complex psychological knowledge, beliefs, and intentions to the players, which has no introspective support and that is not necessary for the account. (See Binmore 2009 for effective criticism; see also Aumann and Brandenburger 1995). Despite his virtue for pursuing game theoretic semantics, Lewis in effect misses one of the main benefits of the program, perhaps its main philosophical payoff, or its main philosophical beauty, which is just how little, *psychologically*, needs to be in place in order to establish linguistic conventions. Wittgenstein was closer to the mark when he emphasized people's lack of reasons for grasping and following linguistic (or other) rules (e.g. Wittgenstein 1958). (Likewise, the handshake is an institution that has evolved culturally without anything like common knowledge). There is mutual convergence on strategies because some strategies are more successful than others, which leads to convergence on certain conventions as solution to signaling coordination problems; but no player has to work out the solution. Those who follow linguistic rules do so automatically, yet not randomly. As Wittgenstein said, they do so 'blindly', without reasons, yet in a way that is nevertheless sensitive to the rule. They go on correctly, and it is no accident that they do—since that is how they have been trained—but they go on without reasons and without beliefs about the contents of other minds as the basis of what they do.

In the evolutionary game theoretic account, the players are assumed to have shared benefits and costs: there are symmetrical payoffs. Hence there are 'incentives' for coordinating their behavior, in the sense that mutual coordination is beneficial; and, thus, there is an increased probability of behaving similarly next time. There is selection because some behavior is reinforced and some deviations from it are costly. That is learning. However, to achieve this, players need not be *consciously cooperating* in their behavior. Players act, and they adjust what they do in the light of what others do and in the light of how well they achieve other goals; thus they *coordinate* their behavior with others without consciously *cooperating*, where doing that includes having beliefs and intentions about other player's beliefs and intentions.

For evolutionary language game theory, there is no need for Gricean meta–meta-representational complications. In particular, language acquisition can be explained by evolutionary language game theory without that. Evolutionary language game theory is *mathematically* complicated, but it is *psychological* simple. This simplicity in itself is an argument in favor of evolutionary language game theory given that simplicity, in the sense parsimony in what is posited, is a theoretical virtue. The counterargument on behalf of a Gricean might be that a mere possibility claim does not prove how it actually is. It might be argued that actual language use in fact proceeds with Gricean meta–meta-representations, even though it is in some sense possible to proceed without them. Given that counter, the search would be for empirical arguments one way or the other. One argument would be that meta–meta-representation is cognitively costly, and therefore would not tend to be the way language is actually acquired, because there would be biological evolutionary pressure against such cost. On the other hand, William Horton and Susan Brennan argue that the meta–meta-representations they investigate, which underlie reference fixing in conversational contexts, is quick and not taxing (Horton and Brennan 2016). So, the appeal to cost is not decisive. A better argument is to appeal to young children. Meta–meta-representation means deploying mental categories in thoughts that have nested intentional contents. But it looks as if basic first language acquisition precedes such meta–meta-representations, and certainly it precedes

complex meta–meta-representations of the kind that Grice and Griceans have in mind. Children display considerable understanding of language before the age of 1 year old, which they manifest behaviorally even though they cannot yet speak. There is some evidence that children of that age can also engage in meta-representation (Onishi and Baillargeon 2005), and perhaps some chimpanzees can do it too. But the meta–meta-representation of other's thoughts about oneself is generally thought to begin after this time, although the matter has not yet been properly researched. It certainly seems that the complex representation of other's representation of one's own representations, as Grice and Griceans require, comes after one and a half years of age. One would have thought that the higher the order of representations, the more difficult is the mental operation, and thus the later it emerges. (Compare the ability to embed conditionals within conditionals). If so, meta–meta-representation emerges after basic language understanding is under way. Ergo, such meta–meta-representation cannot be a prerequisite for language acquisition.[12]

5. Innocence explains sophistication

Even if Gricean models of linguistic understanding are not over-sophisticated, they leave an explanatory gap, since they leave unexplained how it is that people arrive at their beliefs about what others intend by a symbol, and how they can have intentions with such a rich content, which they intend to communicate to someone. They assume something that needs to be explained. Not only are second-order or third-order intentions or beliefs, or semantic beliefs, not necessary for linguistic meaning, they are also not sufficient. Convergence on basic meaning is achievable without sophisticated complex intentions or beliefs about other's beliefs of the sort assumed by those who are under Grice's influence. Furthermore, convergence is achievable without semantic beliefs (that 'by the word 'p' they mean S'). And convergence is achieved without speakers knowing the axioms of some 'theory of meaning' that ordinary speakers supposedly grasp, where those axioms are claims such as ['p' means S] or ['Snow is white' is true if and only if snow is white]. Competent speakers do not need such sophisticated mental furniture. Furthermore, even where there *is* such sophisticated mental furniture, it is the *consequence* of the possession of linguistic ability, not something from which to explain it. Only when the practical ability exists can people think about meanings, and have beliefs and intentions concerning them. Propositional knowledge of meanings depends on linguistic practical knowledge.

The teacher–pupil scenario differs in this respect, for teacher–pupil coordination does depend on conscious cooperation, and presumably, it depends on the possibility of joint attention between teacher (often a parent) and child.[13] But it is crucial that coordination *can* be achieved given evolutionary feedback ensuring convergence without the shared intentions to converge in linguistic behavior that are present in the teacher–pupil scenario. This is how the *sprach–spiel* is set up—how it arises in the first place. There must be some such account or else there would be nothing for teachers to teach and for pupils to learn. The sophisticated teacher–pupil scenario is parasitic on more basic ways of establishing coordination. The relation between teacher and pupil is similar in *some* respects to a group of two or more players who are peers together converging in behavior in setting up a convention; but there are also interesting differences, as teachers deliberately issue rewards and punishments to pupils for getting it right or wrong, and pupils deliberately modify their behavior as a consequence (Wittgenstein 1958, section 1). The convergence on meanings that goes on in the language learning situation is also evolutionary in a broad sense and it can also be modeled game theoretically. The situation, however, could in principle be different from the mutual convergence scenario, since one person is imparting an established meaning to a novice, as opposed to two or more people together establishing a shared meaning. In fact, there is a debate about the role of *negative* feedback in the teacher–pupil situation (see Marcus 1993; Healey et al. 2018; Clark 2020). This may make a difference to the mathematical modeling of the scenario (see further Steels 2011; Brochhagen, 2018). Moreover, all mature players were novices once, which ought to be factored into the modeling of language evolution and transmission. But it does little harm to focus on the mutual convergence scenario rather than on the teacher–pupil transmission scenario.

It is also true that some sophisticated uses of language, such as metaphor or conversational implicature, *do* require sophisticated psychological propositional knowledge of other people. For example, people who call other people 'chicken' know that others will take this to imply cowardice in a certain speech community (in English, but not Hebrew, for example). This use does depend on beliefs about beliefs. Sophisticated culturally local knowledge of other minds and of language is necessary for such uses. However, such sophisticated uses depend on basic word meaning being in place, which is then used or abused in the sophisticated uses, such as metaphor (see Davidson 1978; Zangwill 2014). Such uses are parasites on a host, which is a basic literal meaning, whereby

'chicken' refers to chickens. And for that, the sparser unsophisticated picture is more plausible. On this view, there is a clear distinction between meaning itself, and further metaphorical or conversational uses of meanings, which are self-conscious actions done with words for which basic linguistic meaning is already in place. Such uses depend on existing meanings. Likewise, Harold Bloom might be right that causation in the case of artistic cultural evolution is particularly self-conscious, and not a matter of impersonal 'influences' (Bloom 1973). But the material that is on hand for literary artists to create their works, like clay for potters, was not fashioned in such a self-conscious way. It is those materials that evolutionary game theory describes. It is true that Shakespeare famously self-consciously coined many new words and added to the English language. But whether a word survives and propagates is likely to be a complex matter of interpersonal negotiation. Even when a political movement deliberately introduces or preserves a language, both self-conscious political decision and cultural game theoretic negotiation are factors in what happens to the language. Rational self-conscious linguistic behavior always depends on more basic unsophisticated processes at work. In this respect, biological evolution is analogous, as it hardly ever involves self-conscious reasoning. One special case where human beings breed animals and another is where eugenics policies have been implemented.

Positing sophisticated semantic intentions and beliefs does not explain how there can be semantic facts to intend and to have beliefs about in the first place. It takes such facts and for granted and does not explain how our intentions and beliefs can have semantic content. Of course, we sometimes ascend levels of sophistication, and we develop semantic intentions and beliefs. It is not that there is no truth at all in sophisticated semantic accounts. But that approach cannot be a general theory of linguistic meaning, although some our linguistic lives is how the sophisticated semantic accounts describe it. In contrast, the evolutionary game theoretic program shows how linguistic meaning arises; it gives an explanation, the only plausible one we have, of how linguistic meaning arises from psychological ingredients, as well as what it is. No other account, so far as I know, has a believable *explanation* of how human beings arrive at meanings of natural language symbols, as opposed to assuming it.

6. Psychological parameters of language game theory

The last section was mostly negative; but we need to know what *is* required, psychologically, for evolutionary game theoretical mathematical structures to explain linguistic phenomena. If Gricean sophistication is not necessary for basic meaning, what psychological realities *are* necessary in order to achieve the stable rest points that are word-world linguistic conventions? It is not plausible that *nothing* mentalistic is involved in basic language convention formation. Linguistic meaning is an *achievement*, and it is achieved by establishing conventions or by conforming to established conventions. The way such conventions are established and transmitted is describable in abstract terms by means of the mathematics of evolutionary game theory; but that mathematical structure models real processes and psychological factors are part of those processes. In the case of language acquisition, *some* of the parameters determining convergence on stable equilibrium points (which is what linguistic conventions are) are psychological and some are not. Nevertheless, the psychological parameters may be relatively sparse; they are psychological states that are about things in the environment, such as bears and berries; they are about other player's behavior; and although they may also be about other player's thoughts, they need only be about what in the world other players connect with their symbols. They need not be about what the other players believe or intend about other player's beliefs and intentions. That is, they need not be about other player's meta-representations.

Consider a very simple everyday nonsignaling example. If two people go in opposite directions in a corridor that is just wide enough for two people, they must coordinate their behavior so as to pass each other without collision; and they do that by converging on strategies so that a stable solution is reached. (There are two solutions!) The players desire to reach the end of the corridor soon and without injury. However, the player's trial and error reasoning about this may be entirely behaviorist. Psychological states need not be attributed to the living human mobile obstacle coming in the other direction. There is nevertheless an adapting of behavior by each person to the obstacle that is the other person until they reach a mutually satisfactory equilibrium, and a 'convention' is established, in a sense—even though the two people may never pass each other in a corridor again. Neither sophisticated Gricean meta–meta-representation nor Lewisean 'common knowledge' are needed. The goal of action is just [getting to the end of the corridor quickly and safely]. There is indeed reasoning and each player's beliefs and desires are in play; but there need be no meta-beliefs about the mind of the other. A stable solution can be achieved without that. Signaling coordination need not be very different from nonsignaling coordination. Players need beliefs and desires about

their own positive and negative payoffs when there is coincidence or divergence in assigning a sign a meaning. But the payoffs need not be conceived by players in meta-representational terms. It can just be [That went well] or [That went badly] given the goals of the signaler. At most, they are second-level representational: that the other player thinks that there is a bear there. That is a meta-representational mental state, but it is not a meta-meta-representational or a meta–meta–meta-representational state as is required on Grice's account. Beliefs and desires about the reference are enough.

It is not being denied that signaling activities include beliefs and desires about others. Suppose that two players are in the throes of setting up a convention to establish that some physical symbol means 'bear'. How to do this? Pairs of players do well when both parties associate similar physical symbols with the same things, and they do worse when they differ in what they think of. In this case, they avoid bears or get eaten by them. In the case of divergence in worldly associations, it is hungry bears that enforce the negative payoff to the players! Hungry bears themselves are a cause of convergence and are part of the mechanism of selection. Psychological states of the players also play a role in convergence, in locating the stable strategy, which all or most players share. But these psychological states need not include second-order intentions or beliefs about beliefs or intentions of the players. Getting it right or wrong about bears, and about the other players thoughts about bears, is enough.

This is a partly psychological description of the feedback mechanism. This diverges from many recent philosophers' excursions into evolutionary language game theory, where the approach is more behavioristic (e.g. Shea et al. 2017). But without the psychological aspect of evolutionary language game theory, there is no account of how linguistic meaning derives from or depends on psychological intentionality. (John Searle was right in outline in Searle 1983). There is too great a dislocation between psychological and linguistic intentionality. Linguistic intentionality would not be grounded even in part in psychological intentionality. Indeed, without an explicitly psychological conception of feedback, we are likely either to assume linguistic intentionality rather than explaining it, or else ignore it. The game theoretical account should aim to give an *explanation* of linguistic facts without presupposing them. Therefore, the task is to navigate between two extremes. One extreme is the behaviorist one just noted, which does not at all seek to ground linguistic facts in psychological facts. The other extreme is an over-sophisticated Gricean meta–meta-representational view of the psychological facts in question. In contrast, the combination of simpler psychological facts plus evolutional game theory puts us on the right path between these two extremes.

One of the beauties of game theory is its theory-neutrality, it can apply to diverse subject matters, psychological and nonpsychological. The task of interpreting language game theory includes specifying the replicators and feedback mechanisms. In the case of human language, that mechanism is at least in part a *psychological* mechanism, although it is one that falls well short of what Grice had in mind, as we have seen. But there are also combinations of these psychological realities in which game theoretic structures are partly realized that embody matrixes, and without which the mathematics would not explain concrete phenomena of language. The philosophical interpretation of language game theory specifies the psychological phenomena, which, together with the abstract mathematics of game theory, serves as the mechanism in virtue of which evolutionary game theory explains how linguistic meaning arises and is constituted. The mechanism of selection, then, is partly psychological and partly mathematical. But the mathematical part corresponds to a set of dispositional causal relations. The mathematics computes the way that multiple strategies interact given various variables. Strategies have psychological reality; they are ways that people pursue their goals, where they take account of other's actions. The interactions between strategies are not psychological facts, but are dispositional causal properties with complex relational properties that are only revealed in a mathematical description.

7. Semantic structure and logic

Let us now turn to consider a certain objection to the whole approach. The objection is that the evolutionary game theoretic approach takes language to be merely a collection of names. (Compare the 'Augustinian' picture of Wittgenstein 1953, section 2.) If so, then it seems that it cannot account for the rich structure of language.

Before we address compositionality and logic, consider first sense and reference. Is the evolutionary game theoretic view entirely referential? What about sense? Are there not senses as well as references, as Gottlob Frege argued (Frege 1982)? The intuitive idea of a sense, as opposed to reference, is of a *way* of thinking or referring to things. We can think about the same thing in different ways. And we can use language to refer to the same thing in different ways, and different linguistic items, may embody different ways of thinking about the same thing. Frege builds on this basis, turning intuitive senses into objects; but that is more controversial, and we can leave that extension to one side here. One

evolutionary language game theory model for embracing the intuitive idea of sense would be that different properties of a thing are associated with different names for it. If so, an evolutionary game theoretic approach could say that different names for the same thing have different associated properties, ones that users of names are aware of. Convergence on the reference of names would then take place by means of properties of the object referred to. Those properties, which understanding a name assumes, in turn, may or may not be thought of in different ways. If they can be, the scenario repeats itself. If not, a 'direct reference' (property-less) account opens up. These would be names that refer to objects or properties without senses that are distinct from their references (Donnellan 1966). Alternatively, two names, such as 'Cicero' and 'Tully' could differ in sense due to being set up in different circumstances, despite having the same reference. This might allow for differences in sense without different associated properties. Either way evolutionary language game theory can model senses, in principle. But there is no denying that there is work to do.

What of the systematicity and generative capacity of language, emphasized by Noam Chomsky (Chomsky 1966) Donald Davidson (Davidson 1982). Furthermore, what of the logical complexity of language? What indeed of the language of logic itself? These might be thought to be major problems for the evolutionary game theory of linguistic meaning. The structure of language is a fundamental feature of all adult human languages and any approach that does not account for it is defective. Those invested in the evolutionary game theoretic approach in theoretical linguistics, as far as I am aware, have not worried much about this central aspect of human language (with the exception of Skyrms 2010, to be discussed below). However, we need to consider what this account can say about the structure of language. I shall make some suggestions about this issue with the modest goal of removing the worry that these problems cannot be addressed. What we eventually need is a mathematically worked out model of linguistic and logical structure, but we do need to remove the objection that structure is impossible to achieve on this approach.

One initial point is that the word-world correspondences established by convergence on conventions apply both to names and to predicates; there are linguistic items that refer to objects and properties. Both names and predicates are general in the sense that a name can be applied and reapplied on many occasions, since an object persists over time, and it can be named at different times; and predicates refer to the same property in different instantiations. Conventionality is inherently general. But once we have names and predicates together with the fundamental idea of predication then we have subject–predicate structure. That yields a certain generality—that if we understand 'Fa' and 'Gb' then we also understand 'Fb' and 'Ga', and our ability to understand complex novel sentences is explainable in terms of grasp of their compositional elements.

However, all that assumes that the notion of *predication* itself can be linguistically expressed. There is a question: how can we establish a linguistic convention that expresses the predication relation? We might say 'bear' and 'hungry' but what would be the evolutionary game theoretic account of the meaning of the 'is' of predication? As a preliminary we can note the obvious point that the issue is not about linguistic mood, or about assertion, or about belief; one may have predication in the context of questions, commands and fictional utterances, and hope. Suppose one lacked a linguistic sign for predication. How might one be forged? There seems to be no problem in principle. For example, one possible convention might be that we say 'bear' and 'hungry' while clicking one's fingers in order to mean that the bear is hungry, but if one does not click one's fingers during the utterance of the two words, it might mean that there is a bear nearby and the *speaker* is hungry. That would be one crude convention. A thousand others are conceivable. We can see that it is only natural for syntactic marks of predication to emerge if a payoff structure would encourage the establishment of conventions for signaling predication. Many languages have no dedicated word, but there are nevertheless syntactic phenomena that convey predication.[14] How would the feedback necessary to establish a convention work with predication? Here we must make the meta-physical assumption that the worldly correlate of predication is the instantiation relation. Instantiation is for predication what dogs are for the word 'dog'. Given that, we can say that there are *facts*, which are standardly complex entities consisting of objects or events instantiating properties, which have causal efficacy (Mellor 1995); and that means that facts, which are partly constituted by instantiation relations, can figure in feedback mechanisms that encourage convergence on conventions for predication. So, the evolution of syntactic marks for prediction is intelligible.

Further structure can be explained if simple logic is available. There is no intractable problem here, given the so-called 'introduction' and 'elimination' rules. Their role in *constituting* logic is controversial, to say the least (Prior 1964; Zangwill 2015; Zangwill forthcoming). But their role in enabling convergence on the meaning of logical words is not so controversial. Suppose that there is some feedback mechanism between

players—perhaps smiles or frowns. Suppose 'p' and 'q' individually are both met with smiles. And suppose that 'p*q' is met with a smile also. Then '*' might stand for either conjunction or disjunction. Now suppose 'p' is met with a smile, and 'q' is met with a frown. And suppose that 'p*q' is met with a frown but 'p#q' is met with a smile. Then '*' may well be conjunction and '#' may well be disjunction. That would be the beginnings of a situation in which one person could convey the meanings of logical constant words to another, and we can see how logical constant words could evolve in a community. This scenario assumes confident assignments of meanings to *non*logical constituents, whereas in practice settling nonlogical and logical vocabulary meanings will proceed in tandem, as Quine argued (Quine 1970). Nevertheless, we can in principle envisage an account of how logical language can emerge, since conventions can evolve for referring to logical relations such as *and* or *or*. The story for negation would not be dissimilar. If 'p' is met with a smile and '! p' with a frown, then '!' may stand for negation. A symbol for negation could thus be taught and a convention established, whereby the concept of negation is communicated. An adequate evolutionary game theoretic account of the language of logical constants remains to be developed. Identity and quantifiers, for example, also need to be addressed. My remarks serve merely to show that such a development is feasible in principle.

This kind of approach to the issue of logical structure contrasts with Brian Skyrms' approach (in Skyrms 2010, chapters 11 and 12). He is one of the few writers who have directly addressed the issue. However, his response is limited to postulating functional or dispositional causal relations between different signaling 'subpersonal systems'. (This is endorsed by Huttegger (2014), and a not dissimilar view is suggested in Steels (2011)). This approach is inadequate because it does not engage with the task of showing how logical concepts and thoughts are expressed by interpersonal signaling systems. We think and linguistically express logically complex thoughts rather than merely have dispositions to move between logically simple thoughts. We can think of London *or* Vienna and can say 'London or Vienna', and only because of that have various dispositions. Without logical thought and talk in which logical concepts figure, we have not captured the logic in our thought and language. We need linguistic conventions specifically for logical words in our language, and not just dispositions. They are no substitute. Therefore, Skyrms, and those who follow him on this topic, have not really made a start on the problem of compositionality. Evolutionary game theory as applied to introduction and elimination rules can make such a start. To appropriate Kant's language, we do not merely reason in accordance with logical constants and predicational structure, we do so out of respect for them– because we understand that logical constants or predication are in play, and that is why we are disposed to infer as we do. A purely 'sob-doxastic, sub-personal, account is insufficient. Understanding logical constants and predication has a psychological reality that explains inferential dispositions, and linguistic conventions can be established on the basis of that understanding. Dispositions cannot be where the explanation ends, as it does on Skyrms' account.

Both predication and logic, then, are in principle available to the evolutionary language game theoretic approach. And that means that the compositional and thus recursive aspect of natural languages are also available. There is, at least, no conflict between the game theoretic approach and natural language compositionality. Evolutionary game theorists in theoretical linguistics may venture explanations of semantic and logical structure.

8. Biological evolution and cultural evolution

The evolutionary game theoretic explanation of language is neither incompatible with, nor does it detract from, Noam Chomsky's case for innate grammar as a necessary condition for the learnability of language (Cartesian Linguistic 1966). Nevertheless, innate grammar does not explain how particular linguistic conventions are established, learned, and transmitted, for example, those of English. The grasp of particular linguistic conventions is not innate universal grammar, à la Chomsky.[15] So, it seems that the evolutionary game theoretic account of language acquisition is not in competition with Chomsky's innate universal grammar. Chomsky's account by itself does not and could not explain how contingent linguistic conventions are formed, learned, and transmitted; for that we need the evolutionary game theoretic mechanisms of cultural evolution. The two approaches dovetail nicely. Some have argued that the learnability of many linguistic phenomena via evolutionary game theoretic processes reduces the force of Chomsky's arguments for innate linguistic knowledge (Steels 2017). However, the arguments from poverty of stimulus and universality of grammar are unaffected by game theoretic learnability of specific languages; and we would expect to see the variety that we see in natural languages, despite the universality of grammar that Chomsky finds.

Irrespective of innate grammar, there are biological evolutionary questions to be asked about our innate

capacity to converge in behavior and establish linguistic institutions by which words have meanings. Presumably, there are positive biological payoffs (fitness) for communicating about the environment within groups. Whatever evolutionary game theoretic mechanisms underlie the *cultural* evolution of language—something explored in this paper—they in turn depend on *biological* implanted innate capacities to enact the mechanisms by which linguistic meanings are generated.

So, these two kinds of evolutionary pressures are not completely independent factors affecting human behavior, which have equal status (cf. Tomasello 2012); they stand in some kind of dependence relation. Care is needed, however, in what we say beyond that about the relation between them.

First, even if cultural evolution requires a biological basis, so that human beings have an innate ability to engage in game theoretic convergence, the two evolutionary mechanisms are distinct. The biological evolutionary account tells us *why* we have the capacity, but the cultural evolutionary game theoretic account tells us *what* the capacity consists in. So, even if there is a biological evolutionary explanation underpinning the cultural evolutionary game theoretic account of language, a distinct cultural evolutionary game theoretic account of word meaning is still needed.

Second, since cultural evolution is distinct from biological evolution, cultural evolution can sometimes act contrary to biological evolution. We must not assume that something that is a product of cultural evolution will necessarily enhance biological fitness. Such a step is unobvious and indeed implausible. Indeed, very many successful cultural products are evolutionarily disastrous. (Science, for example, may well lead to our extinction, given its military applications.)

Third, the distinct mechanisms of cultural and biological evolution mean that game theoretic cultural evolution can in turn affect biological evolution, just as cultural knowledge of cooking food has affected human beings' digestive systems. That is compatible with cultural evolution having its basis in innate capacities implanted by biological evolution. Just as the cultural tradition of cooking has affected by our physical natures, likewise, aspects of the cultural game theoretic processes generating language is likely to have affected innate linguistic capacities (Kirby 2017; Smith 2018). Thus, the cultural and biological evolutions of linguistic phenomena coevolve to an extent. Nevertheless, it remains the case that an individual's innate knowledge is a precondition of that individual's capacity to engage in cultural evolutionary convergence processes.

Someone might argue that if communication is so advantageous, surely specific word-world correlations should be innate, not just the capacity to converge on word-world correlations. For example, we could be born with the inclination to connect the sounds corresponding to the English word 'dog' with dogs. So, there would be no room for cultural evolution. One answer to this point is that linguistic conventions are shared among members of a cooperating group, only where there are symmetrical payoffs for those members—that is, where what is good for one is good for the others and what is bad for one is bad for others, unlike in a prisoner's dilemma where the players have competing interests. However, it is often important that some people do *not* understand the language. If word meanings themselves were innate, not just the capacity to converge in meaning and reach an evolutionary stable strategy concerning word-world correlations, then communication would be universal. But that would negate one of the main functions of language, which is to *exclude* as well as to include. It is sometimes said that the function of language is *communication*, but that is naively one-sided. For it is a function of language *both* to facilitate communication among insiders and also to de-facilitate communication with outsiders (see, e.g. Richerson et al. 2016). For this reason, even if some word meanings were innate, there would be pressure for nonuniversal languages to evolve culturally.[16] Furthermore, irrespective of that pressure, linguistic diversity is likely to be a consequence of the cultural evolution of linguistic institutions, which allows variation by random drift in the absence of selection to preserve only some institutions. Cultural evolutionary processes would yield a degree of diversity, even if some meanings were given innately. Hence, biological evolution does not squeeze out cultural evolution.

Cultural and biological evolutionary processes are distinct but interrelated processes.

9. Final comments

The positive case for the cultural evolutionary game theoretic approach to language is, first, that there is no other serious competitor.[17] No other theory has anything approaching an explanation of what linguistic meaning is and how it arises from human interaction. Second, the Gricean phenomena of intentionally using language to generate beliefs by the recognition of that intention, and the existence of mental states with specifically semantic content, while not completely explained by the evolutionary game theoretic account, would be impossible without it, since there would be nothing to communicate, and

there would be nothing to be the intentional contents of semantic beliefs and intentions. Furthermore, the account has testable consequences, as mentioned in the introduction. My brief here has been to give a philosophical interpretation of the evolutionary game theoretical approach to language such that its applicability can be explained, its philosophical benefits appreciated, and worries over compositionality removed.

Many aspects of language remain to be explored with the evolutionary game theoretic approach. For example: (1) the teacher–pupil learning scenario and the mutual convergence scenario need to be integrated with each other and with empirical evidence concerning children's language acquisition; (2) the appropriation of literal meanings in nonliteral uses, such as metaphor, needs to be understood in line with the evolutionary game theoretic approach (Kao et al. 2014; Zangwill 2014); (3) the interpersonal transmission of meanings in Kripke causal-chain cases needs to be modeled (Kripke 1980); (4) the structure of language, especially predication and logic, needs to be part of the whole approach, not just an add-on; (5) a sense/reference distinction needs to be modeled; and (6) nonreferential expressive language needs interpreting (for some first moves see Zangwill 2018). There is work to do. Nevertheless, the evolutionary game theoretic approach seems not only to be philosophically well-grounded, but also well-placed to extend its basic understanding of referential linguistic meaning to many familiar aspects of our linguistic life.

I would like to wrap up this paper by commenting on two matters—first, on the impact of the above defense of evolutionary language game theory on some venerable issues in twentieth century philosophy, and, second, on the kind of interdisciplinarity in play.

First, it might be argued that the evolutionary game theoretic account impacts on central themes of early Twentieth Century Philosophy, such as the nature of logic, necessity and *a priori* knowledge. (See W. V. O. Quine's introduction to Lewis 1969). However, an understanding of linguistic convention along evolutionary game theoretic lines would not by itself make plausible a number of typically verificationist ideas: that logic, necessity, or *a priori* knowledge can be understood in terms as 'analytic truths', where those are thought of as truths that hold by linguistic convention (see, e.g. Ayer 1936; and see also Friedman 1999). There remain grave obstacles to the idea that logic and necessity are matters of convention or that *a priori* knowledge is knowledge of, or by, linguistic convention.[18] One reason is that the existence of conventions may embody only *practical* knowledge on the part of those who participate in the conventions; and there is no reason to think that this practical knowledge is constituted in part by propositional knowledge of meanings. If not, the existence of linguistic conventions is not a solid basis for an interesting theory of analytic truths, one that might be relevant to central philosophical issues over logic, necessity, and *a priori* knowledge. But even if following linguistic conventions requires theoretical knowledge, it would not help. That there are conventions that constitute word meanings, which can be game theoretically understood in a way that involves propositional knowledge, is many miles away from any conventionalist doctrine about logic, necessity and *a priori* knowledge, whereby these are constituted by these conventions. That remains about as plausible as the idea that the laws of physics are constituted by human conventions. There is a large intuitive gap there. Even if some surprising revisionary conventionalism about logic or necessity were to be defended, it would take far more than an understanding of convention, but an understanding of how conventions could possibly constitute logic or necessity, when they are intuitively so utterly different. (There were logical and necessary facts long before human beings evolved and forged conventions). Perhaps, that intuitive gap can somehow be traversed. But traversing the gap will take a lot more than evolutionary game theory.

Second, and last, I want to return to comment on the kind interdisciplinarity that is play, which, I believe, also applies to quite a few other areas where there is an interface between philosophy and other disciplines, whether sciences or humanities. ('Philosophy' is intended in an institutional way). What I have pursued here differs somewhat from what some philosophers who have an interest in the evolutionary game theoretic approach to language have sought to do. That approach recapitulates and to some extent tries to build on the technical account of reference that the signaling account delivered in the 1980s. There is *some* point in this, but only a limited point, in my view. It is, of course, good to raise awareness of the results of one discipline across disciplinary borders. Meanwhile, however, there are tasks for philosophers that may be neglected on the recapitulation approach: first, the mathematical formalisms need interpretation, and adding more formalisms cannot provide that interpretation; second, the main philosophical benefits of such an account need spelling out so that they can be securely archived; and, third, the extension of the program to under-explored linguistic phenomena needs to be signposted if not pursued. In all this, philosophers should work *together* with mathematical game theorists in a useful way, without each partner thinking that they are doing what the other is doing. In particular, the philosophical interpretation must address

mechanism issues, which are not a matter of formal–technical game theoretic mathematics. Evolutionary game theory as a mathematical discipline is one thing, and its application to actual linguistic phenomena, is another, as Rubinstein rightly pointed out. The application of a branch of mathematics to an empirically tractable phenomenon is something that needs to be understood, and that application and understanding lies beyond the mathematics itself. That is for philosophy. Fruitful interdisciplinarity, in my view, means division of labor, rather than each side mimicking what the other side does. A clarity of roles fares better than a muddle in the middle.[19]

Conflict of interest statement. None declared.

Notes

1. David Lewis pioneered its application in the philosophy of language (Lewis 1969). But this was not followed by a flood of interest among philosophers.
2. See Mark Colyvan's elegant exposition of general issues about mathematical explanation of physical phenomena in Colyvan (2001, chapter 3).
3. I ignore the possible existence of 'mentalese' (Fodor 1975).
4. In standard evolutionary models, Nash equilibria are rest points, where everyone loses by defection; but they are not necessarily *stable* rest points. (See Smith 1982; Hofbauer and Sigmund 1998).
5. There has been some recent work by philosophers working on the applications of game theory to language, but they do not address the interpretation and explanatory issues that are the focus of this article. See for example, Zollman 2011; Wagner 2012; Shea et al. 2018. I am not suggesting that these works are inadequate for bypassing by these questions, just that this paper attempts something different in kind.
6. A useful notion of 'mechanism' is not merely a minimal one where a mechanism is constituted by a system of causally interacting elements, since any scattered collection of particles will satisfy that requirement. Instead, elements are taken to constitute a mechanism when what is constituted has certain dynamically stable properties. This makes for an ontological and not merely pragmatic conception of mechanism; contrast, footnote 26.
7. For discussion, see Acerbi and Mesoudi (2015). I am sympathetic to John Dupre's process ontology for biology (Dupre 2013), which should be extended to cultural artifacts.
8. This assumes that negative and positive causal contributions can be separated (see Zangwill 2011).
9. Compare Ruth Millikan's insightful remarks on teleological explanations, Millikan 2002; see also Guest and Martin 2020.
10. Ray Buchanan casts the Gricean requirements in terms of "aims, goals and wants" rather than intentions, but this makes no difference for the issues we are concerned with (Buchanan 2018).
11. It is no accident that Wittgenstein claims, in the preface to Wittgenstein (1953), that an economist was his greatest inspiration and source of criticism: Piero Sraffa. Commentators and followers of Wittgenstein have mostly overlooked this economic approach to his writings on *sprachspiel*. For a notable exception, see Englemann (2013).
12. Richard Moore has attempted to rescue a stripped-down 'Gricean' view that lacks the full-scale meta-meta-representational commitments of Grice's own view (Moore 2017a,b, see also Sperber 2000). But Moore's Gricean view still involves language users having meta-meta-representation, because speakers are said to intend that hearers recognize the speaker's intentions (Moore 2017: 305). (He thinks that nonhuman animals can do that too.) In contrast, evolutionary language game theory requires only meta-representations whereby speakers represent hearers as behaving in various ways or as referring to things, not that speakers represent hearers having psychological states directed to the speaker's intentions. Moore is concerned to make room for the idea that language use can stimulate sophisticated cognitive development. But this is retained on the evolutionary language game-theoretic view. Some somewhat sophisticated meta-representational cognitive abilities are a precondition for game-theoretic language acquisition and use; but language acquisition and use can stimulate further cognitive sophistication, both ontogenetically and phylogenetically, generating more orders of meta-representation. Of course, there may be something we might still call 'Gricean communication', where that only involves mutual knowledge and higher-order representations. And that can be an ability that plays an important role in stimulating later cognitive sophistication. But that kind of Gricean communication does not amount to linguistic communication. (Note that Moore explicitly builds in that the Gricean communicators

intend something with 'an utterance', which assumes that a linguistic institution is in place (Moore op. cit.)).

13. There is some cultural variation in this but not such as to affect the point. See Johnston and Wong 2002.
14. Kirby et al. (2015) give a cultural evolutionary account of compositionality.
15. Perhaps Chomsky's singular phrase '*the* language acquisition device' is problematic.
16. It is harder to *kill* those who speak one's language than it is to kill those who do not (see Browning 1998: 153) since there is a presumed sense of community among common language users.
17. Biological evolutionary accounts of our general linguistic capacities are not in competition with evolutionary game theory as applied to the cultural evolution of specific linguistic conventions.
18. Simon Huttegger has suggested that the evolutionary game-theoretic approach to linguistic convention provides succor for conventionalists about logic who are still reeling from W. V. O. Quine's arguments (Huttegger 2007, 2014; Quine 1936). Quine's most basic argument was that we need logic to follow conventions and that conventions themselves are logically constituted (Zangwill forthcoming). Nothing in what Huttegger says addresses these worries.
19. Many thanks to Christina Pawlowitch for sparking my interest in this approach, for many discussions on and around the topic, and for comments on an early draft. I am also grateful to a referee for this journal who wrote an extensive and helpful commentary, and who put me to work directing me to the empirical literature. Thanks also to Wolfram Hinzen for some helpful challenges. Versions of this paper were given at the Rationality Centre at the Hebrew University of Jerusalem, the University of Helsinki and Tokyo University. The conception of interdisciplinarity expressed here was forged during my years on the steering committee of the Music and Philosophy Study Group, as well as my observation of contemporary philosophy of mind.

References

Acerbi, Alberto and Alex Mesoudi (2015) 'If we are All Cultural Darwinians What's the Fuss about? Clarifying Recent Disagreements in the Field of Cultural Evolution', *Biology and Philosophy*, 30: 481–503.

Argiento, Raffaele et al. (2009) 'Learning to Signal: Analysis of a Micro-Level Reinforcement Model', *Stochastic Processes and Their Applications*, 119: 373–90.

Aumann, Robert and Adam Brandenburger (1995) 'Epistemic Conditions for Nash Equilibrium', *Econometrica*, 63: 1161–80.

Axelrod, Robert (1984) *The Evolution of Cooperation*. New York: Basic Books.

Ayer, Alfred (1936) *Language, Truth and Logic*. London: Dover.

Binmore, Ken (2009) 'Do Conventions Need to Be Common Knowledge?', *Topoi*, 27/1: 17–27.

Brochhagen, Thomas, Michael Franke, and Robert van Rooij (2018) 'Coevolution of Lexical Meaning and Pragmatic Use', *Cognitive Science*, 42/8: 2757–89.

Browning, Christopher (1998) *Ordinary Men*, 2nd edn. New York: HarperCollins.

Buchanan, Ray (2018) 'Intention and the Basis of Meaning', *Ergo 5*, 629–44.

Chomsky, Noam (1966) *Cartesian Linguistics: A Chapter in the History of Rationalist Thought*. New York: Harper & Row.

Clark, Eve (2020) 'Conversational Repair and the Acquisition of Language', *Discourse Processes 57*, 441–459.

Colyvan, Mark (2001) *The Indispensibility of Mathematics*. Oxford: Oxford University Press.

Davidson, Donald, (1978) 'What Metaphors Mean', *Critical Inquiry*, 5: 31–47 (Reprinted in Davidson 1982).

——— (1982) *Inquiries into Truth and Interpretation*. Oxford: Clarendon Press.

Dawkins, Richard (1976) *The Selfish Gene*. Oxford: Oxford University Press.

——— (1986) *The Blind Watchmaker*. Harmonsworth: Pengiun.

Donnellan, Keither (1966) 'Reference and Definite Descriptions', *The Philosophical Review*, 75: 281–304.

Dupre, John (2017) 'The Metaphysics of Evolution', *Royal Society Interface Focus*, 7: 1–9.

——— (2013) 'Mechanism and Causation in Biology', *Proceedings of the Aristotelian Society Supplementary*, LXXXVII: 19–37.

Englemann, Mauro (2013) 'Wittgenstein's 'Most Fruitful Ideas' and Sraffa', *Philosophical Investigations*, 36: 155–78.

Fodor, Jerry (1975) *The Language of Thought*. New York: Thomas Y. Crowell.

Friedman, Michael (1999) *Reconsidering Logical Positivism*. Cambridge: Cambridge University Press.

Grice, Paul (1989) *Studies in the Way of Words*. Cambridge MA: Harvard University Press.

Guest, Olivia and Andrea Martin (forthcoming) 'How Computational Modeling Can Force Theory Building in Psychological Science', Perspectives on Psychological Science.

Healey, Patrick, et al. (2018) 'Running Repairs: Coordinating Meaning in –Dialogue', *Topics in Cognitive Science*, 10/2: 367–88.

Hofbauer, Josef and Simon Huttegger (2008) 'Feasibility of Communication in Binary Signaling Games', *Journal of Theoretical Biology*, 254: 843–49.

—— and Karl Sigmund (1998) *Evolutionary Games and Population Dynamics*. Cambridge: Cambridge University Press.

Horton, William and Susan Brennan (2016) 'The Role of Metarepresentation in the Production and Resolution of Referring Expressions', *Frontiers in Psychology*, 7:1111.

Huttegger, Simon (2007) 'Evolution and the Explanation of Meaning', *Philosophy of Science*, 74: 1–17.

—— (2014) 'How Much Rationality Do we Need to Explain Conventions?', *Philosophy Compass*, 9: 11–21.

Kao, Justine, Leon Bergen, and Noah Goodman (2014) 'Formalizing the Pragmatics of Metaphorical Understanding' in *Proceedings of the Annual Meeting of the Cognitive Science Society*, 36.

Kirby, Simon (2017) 'Culture and Biology in the Origins of Linguistic Structure', *Psychonomic Bulletin and Review*, 24/1: 118–37.

Kirby, Simon et al. (2015) 'Compression and Communication in the Cultural Evolution of Linguistic Structure', *Cognition*, 141: 87–102.

Kripke, Saul (1980) *Naming and Necessity*. Oxford: Oxford University Press.

Lewis, David (1969) *Convention*. Oxford: Blackwell.

Lowe, Jonathan (1996) *Subjects of Experience*. Cambridge: Cambridge University Press.

Machamer, Peter, Lindley Darden, and Carl Craver (2000) 'Thinking about Mechanism', *Philosophy of Science*, 67: 1–25.

Marcus, Gary (1993) 'Negative Evidence in Language Acquisition', *Cognition*, 46: 53–85.

Maynard-Smith, John (1982) *Evolution and Theory of Games*. Cambridge: Cambridge University Press.

McGinn, Colin (1984) *Wittgenstein on Meaning*. Oxford: Blackwell.

Mellor, David (1995) *The Facts of Causation*. London: Routledge.

Millikan, Ruth (2002) 'Biofunctions: Two Paradigms' in Ariew, Andre, Robert Cummins, and Mark Perlman (eds.) *Functions: New Readings in the Philosophy of Psychology and Biology*. Oxford: Oxford University Pres: 113–143.

—— (2005) *Language: A Biological Model*. Oxford: Oxford University Press.

Morgenstern, Oskar and John von Neumann (1944) *Theory of Games and Economic Behavior*. Princeton: Princeton University Press.

Nash, John (1950) 'Equilibrium Points in n-Person Games', *Proceedings of the National Academy of Sciences*, 36/1: 48–49.

Nowak, Martin and David Krakauer (1999) 'The Evolution of Language', *Proceedings of the National Academy of Sciences of the United States of America*, 96/14: 8028–33.

Nowak, Martin, Joshua Plotkin, and David Krakauer (1999) 'The Evolutionary Language Game', *Journal of Theoretical Biology*, 200: 147–62.

Onishi, Kristine and Renee Baillargeon (2005) 'Do 15-Month-Old Infants Understand False Beliefs?', *Science*, 308: 255–58.

Pawlowitsch, Christina (2008) 'Why Evolution Does Not Always Lead to an Optimal Signaling System', *Games and Economic Behavior*, 63: 203–26.

Prior, Arthur (1964) '*Conjunction* and Contonktion *Revisited*', *Analysis*, 24/6: 191–95

Quine, Willard Van Orman (1936) 'Truth by Convention'. *Reprinted in the Ways of Paradox*, 2nd edn. Cambridge: Harvard University Press, 1976.

—— (1960) *Word and Object*. Cambridge, MA: Harvard University Press.

—— (1970) *Philosophy of Logic*. Cambridge, MA: Harvard University Press.

Richerson, Peter et al. (2016) 'Cultural Group Selection Plays an Essential Role in Explaining Human Cooperation: A Sketch of the Evidence', *Behavioral and Brain Sciences*, 39: 30.

Rubinstein, Ariel (1991) 'Comments on the Interpretation of Game Theory', *Econometrica*, 59: 909–24.

Searle, John (1983) *Intentionality*. Cambridge: Cambridge University Press.

Shea, Nick, Peter Godfrey-Smith, and Rosa Cao (2017) 'Content in Simple Signaling Systems', *British Journal for the Philosophy of Science 38: 1009-1035*.

Shoemaker, Sydney (1981) 'Some Varieties of Functionalism', *Philosophical Topics*, 12/1: 83–118.

Schiffer, Stephen 1972: *Meaning*. Oxford: Clarendon Press.

Skyrms, Brian (1992) 'Chaos and the Explanatory Significance of Equilibrium: Strange Attractors in Evolutionary Game Dynamics', in *Proceedings of the Biennial Meeting of the Philosophy of Science Association, Volume Two: Symposia and Invited Papers*: 374–94.

Skyrms, Brian (2010) *Signals*. Oxford: Oxford University Press.

Smith, Kenny (2018) 'How Culture and Biology Interact to Shape Language and the Language Faculty', *Topics in Cognitive Science*, 12: 690–712.

Sperber, Dan (2000) *Metarepresentations: A Multidisciplinary Perspective*. Oxford: Oxford University Press.

—— and Deirdre Wilson (1986) *Relevance: Communication and Cognition*, 1st edn. Oxford: Blackwell.

Steels, Luc (2011) 'Modeling the Cultural Evolution of Language', *Physics of Life Reviews*, 8: 339–56.

—— (2017) 'Human Language is a Culturally Evolving System', *Psychonomic Bulletin & Review*, 24: 190–93.

Tomasello, Michael et al. (2012) 'Two Steps in the Evolution of Human Cooperation', *Current Anthropology*, 53: 673–75.

Wagner, Elliott O. (2012) 'Deterministic Chaos and the Evolution of Meaning', *British Journal for the Philosophy of Science*, 63: 547–75.

Wiggins, David (1997) 'Meaning and Truth Conditions: From Frege's Grand Design to Davidson's'. In: Hale, Bob, Wright, Crispin (eds.) *A Companion to the Philosophy of Language*. Oxford: Blackwell: 3–28.

Wittgenstein, Ludwig (1953) *Philosophical Investigations*. Oxford: Blackwell.

——(1958) *Blue and Brown Books*. Oxford: Blackwell.

—— (1993) 'Remarks on Frazer's *Golden Bough*'. In: *Philosophical Occasions*. Indianapolis: Hackett.

Woodward, James (2013) 'Mechanistic Explanation: Its Scope and Limits', *Proceedings of the Aristotelian Society Supplementary Volume*, LXXXVII: 38–65.

Zangwill, Nick (2011) 'Negative Properties', *Noûs*, 45: 528–56.
—— (2014) 'Metaphor as Appropriation', *Philosophy and Literature*, 38: 142–52.
—— (2015) 'Logic as Metaphysics', *Journal of Philosophy*, CXII: 517–50.
—— (2018) 'The Yummy and the Yucky', *Monist*, 101: 294–308.
—— (forthcoming) 'Against Logical Inferentialism: Stipulation, Rules and the Mushroom Omelette Problem', *Logique et Analyse*.
Zollman, Kevin (2011) 'Separating Directives and Assertions Using Lewis Signaling Games', *Journal of Philosophy*, 108: 158–69.