Philosophical Psychology
Vol. 23, No. 1, February 2010, 59–74

Routledge
Taylor & Francis Group

Can computational simulations of language emergence support a 'use' theory of meaning?

Whit Schonbein

Some researchers claim that simulations of the emergence of communication in populations of autonomous agents provide empirical support for 'use' theories of meaning. I argue that this claim faces at least two major challenges. First, the empirical adequacy of such simulations must be justified, or the inference from simulation results to real-world linguistic behavior must be dropped; and second, the proffered simulations are in fact compatible with all of the competing theories of meaning surveyed, suggesting that theories of meaning are not the kinds of theories for which simulations provide evidence. To conclude, I consider what impact this has on the project of developing a naturalized theory of language.

Keywords: Computational Simulation; Naturalism; Theories of Meaning

1. Introduction

The use of computational simulations is firmly entrenched in contemporary research on the emergence of language (Cangelosi & Parisi, 2002). Their utility derives from their potential to play significant epistemological roles not readily or easily achievable by other methods. For example, Steels and Belpame (2005), speaking of the value of computational simulations ("theoretical models," in their terminology), list the following virtues:

> Theoretical models make a particular view explicit, thereby making it easier to structure the debate for or against a certain position. Theoretical models bring out the hidden assumptions of an approach, particularly with respect to the cognitive mechanisms that are required and the information they need. Moreover, they help to assess the plausibility of certain assumptions—for example, with respect to the time needed to acquire categories or propagate word-meaning pairs in a

Whit Schonbein is an Assistant Professor at the College of Charleston.
Correspondence to: Whit Schonbein, Department of Philosophy, College of Charleston, 14 Glebe Street, Charleston, SC 29424, United States. Email: schonbeinw@cofc.edu

ISSN 0951-5089 (print)/ISSN 1465-394X (online)/10/010059-16 © 2010 Taylor & Francis
DOI: 10.1080/09515080903532258

> population. Finally, theoretical models may suggest new experiments for empirical data collection. (p. 472)[1]

However, such benefits are undermined when simulations are used for tasks for which they are in principle not applicable. One such task, I contend, is that of providing evidence for or against theories of meaning, in the philosophical sense of the phrase. Theories of meaning, I argue, are simply not the kinds of theories for which simulation studies are relevant.

Some simulation studies of language emergence invite interpretation in terms of 'use' theories of meaning. Examples include those simulations appearing in Dyer (1995), Hutchins and Hazelhurst (1991), and Steels and Vogt (1997). Some of these authors clearly take inspiration from use theories of meaning, so it is natural to interpret the simulations from that perspective, but the authors do not explicitly claim that the models provide support for use theories. In a more recent simulation study, Grimm, St. Denis, and Kikalis (2004; henceforth GSK) appear to provide an explicit instance of this reasoning. The argument of this paper is thus directed primarily at GSK's formulation, but extends to parallel arguments based on other simulation studies as well: Theories of meaning are not the kinds of theories that can be tested by appeals to simulations of language emergence.

I begin by presenting an interpretation of GSK's argument, including a brief description of their simulation study, highlighting the features it shares with other simulations. In the section following I describe two objections, exploring the second at length. The first is that the simulations do not satisfy conditions on epistemological adequacy; and the second is that the simulations are compatible with all purportedly incompatible theories of meaning, and consequently do not provide support to just one. I conclude that, while there may exist a suitable response to the first objection, the fact that all surveyed theories of meaning are compatible suggests that theories of meaning in general are resistant to empirical investigation through simulation studies. To conclude the paper, I argue that, first impressions notwithstanding, this presents no threat to the project of providing a naturalized theory of language.

2. Simulations of Language Emergence and Theories of Meaning

Some well-known theories of language hypothesize that word meanings are mind-independent objects (e.g., referents or senses) or mental entities (e.g., ideas) to which linguistically competent agents are related (e.g., by 'grasping' the meaning). In contrast to these theories—all of which identify some type of 'thing' as *the* meaning for a word—'use' theories deny that there are any such things. Instead, the meaning of a word is determined by the way it is used in a linguistic community. So, rather than deriving their meanings from standing in some relationship to an entity, property, mental state, etc., words are meaningful by virtue of residing in a network of implicit rules or conventions governing their use in a community.[2]

Over the past decade (at least) there have appeared a number of research programs involving the computational simulation of language emergence that draw inspiration from 'use' theories. Examples include work by Luc Steels (e.g., Steels & Vogt, 1997; see also Dyer, 1995; Hutchins & Hazelhurst, 1991). While these researchers make use of a variety of simulation methods—e.g., embodied (robots) versus fully simulated agents, lookup tables versus artificial neural networks, or individual versus evolutionary learning—they are structurally similar insofar as they do not posit, as part of the starting conditions of the simulation, that some specific entities function as 'the meanings' of the words comprising the emergent shared lexicon. That is, the criterion for successful communication is not taken to be 'grasping the same meanings' in the sense of accessing the same entity, but rather coordinated behavior.[3] In general, the methodology behind these experiments is to populate an environment with a finite number of individual agents, and then provide pressure—through artificial evolution, competitively solving a problem, or training agents to behave more like their neighbors—such that, after some number of iterations of the process, a shared lexicon emerges in the sense that members of the population make the same utterances in the same contexts, and behave the same way when perceiving the same utterance. So, for example, we might say that agent *X* communicates successfully to agent *Y* with utterance *U* when *Y* behaves appropriately given that utterance (where what it is to behave appropriately depends on the task demands). Nowhere do shared meanings enter the picture, either as internal 'ideas' or external abstract objects, and in this sense the prima facie affinity between the general structure of these simulations and 'use' theories of meaning is clear.

This parallel has more recently been marshaled into an explicit argument in support of 'use' theories of meaning. In their (2004) paper, Patrick Grimm, Paul St. Denis, and Trina Kokalis contrast 'relational' theories of meaning—according to which meanings are a sort of shared entity—with the results of their own series of computational simulations of communication emergence. GSK's simulations consist of a two-dimensional toroidal grid, where each location contains an immobile agent modeled as a neural network. Floating randomly across this grid are bits of food (which persist through being consumed) and predators. At each time step, each agent will (i) either hide, open its mouth, or do nothing, and (ii) possibly emit an utterance. Each agent has a point score subject to the following modifications: the score is deducted when the agent opens their mouth, hides, emits a noise, or when a predator lands on the agent's location when they are not hidden. The score is increased when food lands on the agent's location and their mouth is open. The agents react to the inputs they are given, and these inputs include (i) the presence or absence of food on their location, (ii) the presence or absence of a predator on their location, and (iii) any sounds emitted by itself or one of its eight neighbors.

A simulation begins with an array of randomly configured agents. After running it for 100 iterations, the following procedure is used to modify the network weights: for each agent, select the neighbor with the highest point score above the agent's (if one exists), and partially train the agent on a subset of the neighbor's input-output profile. The overall trend enforced by this training regimen is (in most cases) to push the average score of the population of agents upwards: the best strategy for maximizing an agent's score is for that agent to use information contained in the utterances of its neighbors to guide its behavior for the next time step. For instance, all things being equal, if a neighbor signals that food is present the agent should open its mouth on the next time step in the hope that the food's random walk will take it there. Consequently, agents that are able to take advantage of neighbor's utterances in cases where those utterances are reliable indicators will have higher scores. Since agents are trained on the input-output profiles of neighbors, the regimen also pushes the agents towards a shared lexicon. In all cases, communities of signaling agents emerge: One sound is used to signal the presence of food, causing neighbors to open their mouths, while a distinct sound is used to signal predators, causing neighbors to hide.[4]

As in the simulations mentioned previously, GSK's experiments make no obvious appeal to anything resembling a shared meaning, either external (e.g., senses or referents) or internal (e.g., ideas). This can be appreciated by noting that the criteria for communicative success in these simulations is not the access of shared meanings, but rather successful behavioral coordination—the community is coordinated in that they all use the same words to provoke the same behavior in the same contexts. GSK hold that their simulations consequently demonstrate something about the nature of linguistic meaning. They write:

> On the view we try to model here, a grasp of what meaning and information are will come not by looking for the right relations involving the right kind of object—meanings, for example—but by attention to the coordinated interaction of agents in a community. In practical terms, the measure of meaningful communication and information transfer will be functional coordination alone. The understanding we seek may thus come with an understanding of the development of patterns of functional communication within a community, but without our ever being able to identify a particular relation as the 'meaning' or 'information' relation, and without being able to identify a particular object—concrete, ideational, or abstract—as the thing that is the 'meaning' of a particular term . . . (GSK, 2004, p. 46)

Indeed, GSK suggest that their simulations provide some form of *evidence* for 'use' theories of meaning when they say, "the surprising results that appear in these models . . . offer important *support* [italics added] for an approach to both meaning and information in terms of use" (GSK, 2004, p. 46). Similarly they claim, "for the patterns of meaning and information transfer that seem to appear in the models offered here . . . ideational theories *are significantly less plausible* [italics added] than a theory written directly in terms of pattern of use," (GSK, 2004, p. 62). So, in these passages it is GSK's contention that the data delivered by their simulations increases the likelihood (or our reasonable expectation) that some form of 'use' theory is true of natural language.

The following justification for the second of the quoted claims is given:

> What arises in a community [of simulated agents] is a pattern of coordinated behavior, but in evolving from an initially randomized array of neural nets that

> coordinated behavior need not be built on any uniform understructure in the nets themselves. There is thus no guarantee of matching internal representations in any clear sense, no guarantee of matching internal 'meanings' or internal 'pieces of information' that are transferred, and no need for internal matches in the origin and maintenance of patterns of communication or information across a community. (GSK, 2004, p. 63)

In other words, according to an ideational theory, all agents who express the same meaning share the same idea, where an idea is an internal state. The simulated agents, however, can express the same meaning (as is evidenced in their coordinated linguistic behavior) without shared internal states. Therefore, shared ideas are not essential components of word meanings. Similar reasoning applies for other relational theories of meaning. The only remaining option is that the ways words are used by the community constitute their meanings.

It is clear that arguments with the same structure can be constructed on the basis of the other simulation studies described above. In general, if a simulation results in a shared system of linguistic communication without relying on some version of the assumption that meanings are 'things' shared between competent members of the language community, it may be taken as support for a 'use' theory of meaning. Since the simulations referred to above satisfy the antecedent condition, then, if GSK's reasoning is sound, they also support 'use' theories of meaning.

Before investigating this reasoning in more detail, it is useful to consider other interpretations of GSK's argument. As I've reconstructed it, the simulations are taken to provide data increasing the likelihood that a use theory is true of natural language (relative to competing theories). An alternative interpretation is that GSK appeal to Occam's razor: it *could* be the case that explaining linguistic behavior requires positing some form of shared meaning, but these simulations show that the phenomenon of interest (or at least a very simplified analog of it) can be realized *without* positing shared meanings. Therefore, since the simpler explanation should be preferred, the simulations favor a 'use' theory. Finally, a third interpretation is that GSK's simulations are simply supposed to serve as proofs-of-concept, thereby dispelling any doubt that 'use' theories can account for the phenomenon in the first place. In this case the models do not render 'use' theories more likely, but are instead viewed as 'how-possible' models, standing alongside other possible accounts (see GSK, 2004, p. 47, for language that suggests this interpretation).

The passages quoted above suggest to me that GSK conceive of their simulations as playing a stronger epistemic role than merely demonstrating the possibility of a 'use' theory. But rather than debate the proper interpretation, I will simply make two points. First, the issue of whether these sorts of models provide support for a theory of meaning is important for understanding the epistemological role of computational simulations in philosophy regardless of how one interprets GSK's particular position. Second, my argument will apply to all of these interpretations: Since (the argument goes) the simulations are consistent with all of the competing theories of meaning, they are not realizations of any particular theory of meaning, and hence they do not show that the phenomenon of interest can be realized without positing shared meanings.

3. Epistemological Adequacy and Underdetermination

As previously described, there are at least two possible responses to a GSK-style argument: the first challenges the epistemological adequacy of the simulations, and the second argues that they are in fact irrelevant to the debate.

The challenge to epistemological adequacy begins with the observation that, as in other mathematical and computational models, the suggestive force of the simulations depends on their dynamics being similar to the dynamics manifested in actual linguistic behavior: It is this similarity that allows the inference from claims about simulation results to claims about real-world linguistic systems. A detractor may thus attempt to argue that the purported similarity is not justified—i.e., the simulations are not epistemologically adequate—and consequently the inference fails: the dynamics of the simulation may be interesting, but without further argument, they fail to imply anything about the dynamics of natural language use.

Work on the epistemology of computational modeling suggests one possible elaboration of this objection. Simulations of language emergence contain many simplifying assumptions that are strictly speaking false. Several authors have noted that such assumptions are not necessarily fatal to the capacity of a model to generate relevant data, and in some cases may even be beneficial (Norton & Suppe, 2001; Weisberg, 2006). However, it is still the case that some method for detecting and correcting errors introduced by those assumptions is required. So, for example, in their discussion of atmospheric modeling, Norton and Suppe (2001) conclude:

> Simulations may embody false or unsubstantiated simplifying assumptions and still detect real effects. One has to determine which aspects of simulations are real effects versus artifacts and tailor claims to the reliability and limitations of the data-model output. (p. 91)

Elaborated in this way, the objection is that simulations of language emergence such as GSKs lack methods for measuring and correcting for artifacts introduced by various simplifying assumptions. Without such assurances, inferences from features of the model to features of the target system—human linguistic behavior—are unwarranted.

This objection is obviously very broad: if it applies to GSK's simulations, then it likely applies to many other computational models in the cognitive sciences. So it is really an issue about the epistemological function of computational models in general. Because this line of inquiry leads to such broad issues outside the scope of this paper, I turn instead to the second potential objection.

As mentioned above, GSK address theories of meaning very generally, distinguishing between 'relational' theories on the one hand, and 'use' theories on the other. The defining feature of a 'relational' theory is that meanings are some sort of entity to which agents are related. Examples of relational theories include purely referential theories (where the meaning of an expression is the object it refers to), ideational theories (where the meaning of an expression is the internal representation it is used to express), and Fregean theories (where the meaning of an expression is a mind-independent, immaterial sense).

Consider first the pure referential theory. The basic idea behind a contemporary referential theory is that there is some form of causal relation between an expression and the things it refers to, and this causal relation is responsible for establishing reference or meaning. One possible relation is *causal co-variation*, where an expression *W* refers to objects of type *A* because tokenings of *A* cause tokenings of *W*, and non-*A*'s do not. This simple referential theory of meaning is consistent with GSK's simulations: in these simulations we have evolved causal co-variations between expressions and predators and food. The expression for predator causally co-varies with the presence of predators and not with the presence of food, and likewise for the expression for food and the presence and absence of food. Since this is a case of reference fixation in accordance with the causal co-variation theory, GSK's simulations can be viewed as implicating a referential theory of meaning.[5]

This primitive causal co-variation account faces well-known difficulties, most notably the 'disjunction problem': suppose some property *P* not in the extension of *W* causes *W* to be tokened. According to the basic causal co-variation account, *P* must therefore be included in the extension of *W*, which conflicts with the initial assumption that *P* is not in its extension (Fodor, 1990).

There are also other well-known problems for referential theories—not limited to co-variation accounts—such as failures of substitution in intensional contexts. For example, while it may be true that George enjoys the music of Bob Dylan, it may be false that he enjoys the music of Robert Zimmerman, regardless of the fact that "Bob Dylan" and "Robert Zimmerman" are coextensive. Consequently, the meaning of those names cannot consist merely in their reference.

Since neither of these challenges to the co-variation account arises within the simulations, a proponent cannot appeal to them as reasons to reject interpreting the simulations as utilizing a referential theory. But the existence of such challenges suggests a potential response: The problem, a simulation proponent may claim, is merely that the simulations are too simplistic. If they were made more realistic, then the disjunction and substitution problems would emerge *within* the simulation, and therefore the simulation would no longer be consistent with (simple) referential theories. In short, a simulation proponent might bet that more realistic models will provide evidence against referential theories insofar as such simulations will exhibit phenomena known to cause difficulties for those theories. The problem is in the complexity of the simulation, not with the project itself.

Let us suppose that a more complicated simulation exhibits behavior interpretable as the disjunction problem. For example, looking at the simulation results, we note that both *A*'s and *B*'s cause instances of *W*, but the extension of *W* contains only *A*'s, and hence a causal co-variation theory fails to apply to the simulation, and the simulation constitutes a counterexample to the co-variation account. The problem with this argument lies in our assumption that the extension of *W* contains only *A*'s, for this is precisely what the primitive co-variation theory denies: The co-variation theory says that *B*'s are included in the extension precisely because they cause tokens of *W*. Interpreting the simulation as exhibiting phenomena instantiating the disjunction problem begs the question against the basic co-variation theory since it presupposes there is more to determining extension than causal co-variation.

Similarly, suppose a more complicated simulation exhibited linguistic behavior interpretable as intensional contexts. Here's how that might work: (i) agents categorize objects through sets of perceptual primitives; (ii) in at least some cases different sets are used to categorize the same object; and (iii) different sets are associated with different terms in the shared lexicon. If we conceive of these sets of primitives as 'descriptions' of the categorized object, then different terms reflect different descriptions. For example, let term T_1, associated with feature set $\{p, q, r\}$, be shorthand for the description 'an object with features *p*, *q*, and *r*', and let T_2, associated with $\{s, t, u\}$, be shorthand for 'an object with features *s*, *t*, and *u*'. So, an agent will utter T_1 when presented with an object that causes set $\{p, q, r\}$ to be tokened, and will utter T_2 when $\{s, t, u\}$ is tokened. Now, assume that T_1 and T_2 refer to the same type of object. Finally, suppose we have reason to interpret utterances by agents in the community as belief-reports about conspecifics (e.g., suppose the task demands of the simulation are to issue such reports). It then follows that, when an agent utters '*B* believes that that T_1 is present', where T_1 is interpreted as shorthand for 'an object with features *p*, *q*, and *r*', we cannot substitute T_2 for T_1 in *B*'s utterance, since, while *B* does believe that there is an object with features *p*, *q*, and *r* present, it might not believe there is an object with features *s*, *t*, and *u* present. Consequently, we conclude that the simulation exhibits failures of substitution, and cannot be interpreted as instantiating a pure referential theory of meaning.

Unfortunately, this argument begs the question against pure referential theories insofar as it assumes the truth-values of the two utterances are different. A pure referential theory predicts that their truth-values are identical precisely because they differ only in a single co-referential term, whereas the opaque interpretation denies this assumption without argument.[6]

The fact that both the disjunction problem and opaque context defenses are question-begging suggests something more general about the role of simulations in evaluating theories of meaning: the adoption of a theory of meaning is *prior* to our interpretation of any simulation results. In the co-variation case, a simulation only exhibits the disjunction problem when we approach the problem of interpreting its results with a non-referential theory already in hand. Similarly, a simulation will exhibit opaque contexts only when we previously decide to interpret features of the simulation as opaque. This is why both attempts at defending the simulations as evidence against a referential theory beg the question—they implicitly assume that referential theories are incorrect.[7]

Next, consider ideational theories. Ideational theories hold that the meaning of a word is the idea it expresses. GSK interpret simulations that rely on syntactically matching internal representations as presupposing an ideational theory of meaning (GSK, 2004, p. 45), so they implicitly understand ideational theories to require matching internal representations as a criterion of communicative success. Their reasoning might be the result of a traditional objection to ideational theories: no two people have the same ideas, and, since meanings are shared, meanings are not ideas. One way to respond to this criticism of the ideational theory is to require that internal representations match. Since GSK's simulated agents realize the same input/output profile with different weight configurations, they do not have matching internal representations, and hence they are not ideational. Therefore, coordinated linguistic behavior does not require an ideational theory.

The strategy in dealing with this argument mirrors the treatment of referential theories: first, we'll see that the simulations are in fact consistent with versions of an ideational theory, and second, attempts to interpret simulation results as problematic beg the question against ideational theories.

A contemporary proponent of an ideational theory may appeal to some other shared feature of internal representations besides their *physical form* to establish that those represents do in fact match. An obvious candidate in the case of GSK's simulations is *functional role*: if the content of an idea is determined by the functional role it plays mediating between inputs and outputs, then perhaps every agent possesses internal states that have the same content. This is a distinct possibility given the simplicity of the simulation agents; yet even if we were to return to GSK's data and discover functional differences within linguistic communities, this would not suffice to settle the issue, because more sophisticated functionalist accounts of mental content are available to address variation across individuals while preserving shared content. So, a functionalist version of an ideational theory is consistent with GSK's simulations.

Furthermore, ideational theories are not necessarily committed to there being shared content in the first place. Taking inspiration from chapter six of Locke's *Essay*, we note that two agents with distinct ideas may nonetheless coordinate behavior, provided that no situations emerge to reveal the lack of internal agreement. For example, suppose Homer represents gold as being a yellow metal, and Bart represents gold as being a heavy, precious element. Because of their different ideas of what constitutes gold, Bart and Homer would draw different inferences about gold if situations requiring such inferences were to arise. But provided they do not, neither will be the wiser, and their behavioral coordination with respect to gold will be preserved. In contrast to the functionalist strategy described above, which attempts to preserve shared content despite behavioral differences, ideational theories adopting this strategy simply concede the point—the agents express different contents with the same word—noting that this is not an issue for their theory. So an ideational theorist can deny that the agents lack matching representations, or she can deny that matching representations are even required.

What sorts of simulation evidence *would* provide support for a use theory over an ideational one? Imagine we have an incredibly complicated simulation that exhibits what appears to be linguistic behavior on par with that of human beings. Given that an ideational theory need not require *formally* matching internal representations, we can't just look at the internal representations of the agents to answer the question.

Furthermore, we cannot just look at the functional roles of internal states, because even if they don't agree, a fully developed functionalist theory of content may nonetheless be able to deliver matching contents. Finally, we might decide to adopt a functional role theory that does not require matching content at all. It seems, then, that there is nothing in the simulation itself that decides the issue.

Instead of appealing to the simulation for evidence, we must appeal to our own linguistic intuitions. For example, one difficulty for any theory that denies that the content must match is guaranteeing deeply held intuitions about the nature of truth: if the word 'meter' (for example) means one thing for agent *A* and something else for *B* (by virtue of expressing different ideas), then in what sense is *A* uttering a *true* sentence when she says, 'There is a rod in Paris one meter long', while *B* asserts something *false* in asserting its negation? Since the meaning of 'meter' differs between *A* and *B*, both of their utterances are true, which (according to intuition) is absurd. Similarly, a fundamental challenge to (individualist) ideational theories depends on our intuitions about what our words would mean if the world were different than it actually is, i.e., appeals to modal intuitions (e.g., Putnam, 1973). In both cases, it is hard to fathom how a simulation could possibly provide versions of these arguments, since any simulation would first have to be interpreted as conforming to the very intuitions that lead to a rejection of the ideational theory in the first place. That is, such appeals would beg the question.

This same concern emerges for Fregean theories, the essence of which is the thesis that meanings are mind-independent, immaterial entities or Senses. Understanding the meaning of a word requires that one stand in a 'grasping' relation to the appropriate Sense. Of all the theories considered, it would seem that Fregean theories are clearly inconsistent with GSK's simulations, for three reasons. In GSK's simulations, (i) there are no simulation elements playing the role of mind-independent Senses; (ii) there is no explicitly represented 'grasping' relation; and (iii) the criterion for successful communication is that linguistic behavior is coordinated, not that agents 'grasp' the same mind-independent entity. Since the agents *can* coordinate linguistic behavior *without* relying on such entities and relations, the simulations provide relative support for a use-theory.

In reply, I suggest a Fregean can argue that the simulations provide an account of the causal mechanisms underlying linguistic behavior, including the mechanisms of grasping Senses, but do not say anything about the nature of meaning. In other words, reason (i) is irrelevant, since a simulation of the causal mechanisms of linguistic behavior requires no explicit representation of Senses; (ii) is false, because grasping is a psychological (brain) state of an agent, and (for independent reasons) the agents must be grasping Senses; and (iii) equivocates on the notion of 'successful communication': Shared understanding requires grasping the same Sense, but causally coordinating behavior does not.

To motivate this position, notice first that it is a reasonable one for a Fregean to adopt. For one thing, the distinction between grasping meanings and coordinating behavior is reflected in traditional Fregean arguments in favor of Senses, none of which attempt to illuminate the mechanisms of linguistic behavior. Rather, they address issues such as the cognitive significance of identity statements involving coextensional terms. Since Fregean theories are not offered as accounts of the mechanisms of linguistic coordination, it stands to reason that they function *alongside* theories of linguistic behavior. Furthermore, causation is typically understood to involve only material entities, properties or events; immaterial entities cannot stand in causal relations with material entities. Given that Senses are immaterial, we shouldn't expect Senses to play a *causal* role in linguistic behavior. Mechanistic accounts of linguistic coordination such as GSK's simulations are causal stories—they trace the causal interactions between agents, their internal states, and the environment such that linguistic behavior is produced. So, again, we should expect that Fregean theories are consistent with GSK's simulations.

To verify this expectation, we must consider where a Fregean theory might have implications for the causal mechanisms of linguistic behavior, for it is here that conflicts between simulation and theory will arise. The point of 'contact' between an agent and a Sense is the process of 'grasping' that sense, which, following Putnam (1973), we assume is a psychological state of an agent. So, for example, I understand the word 'water' by instantiating a psychological state constitutive of my grasping the associated Sense. Assuming further that psychological states are internal states of agents, if a Fregean theory has any implications for the causal organization of the simulated agents, they will arise as constraints on the internal states of agents, i.e., on how agents grasp Senses. So we are looking for situations where the Fregean and simulation accounts disagree over some aspect of this point of contact. The two obvious candidates for disagreement are the following claims to which the Fregean appears to be committed: (a) if two agents instantiate type-identical psychological states, they grasp the same Sense, and (b) if two agents grasp the same Sense, they are in type-identical psychological states. The idea, then, is that a simulation could test these claims by showing that two agents instantiating identical states can nonetheless grasp different meanings, or that two agents grasping the same meaning may be in different psychological states.

The standard argument against (a) is Putnam's familiar Twin Earth thought experiment (Putnam, 1973). Briefly, suppose I understand 'water' by instantiating a brain state that grasps its Sense. Now throw me through a wormhole to Twin Earth, which is identical to Earth except the clear, odorless, colorless liquid found in lakes is not H_2O but rather of a distinct atomic structure, XYZ. Pointing at a lake while tokening the water-grasping brain state I say, 'That's water'. But my claim would be false because the Sense I express with 'water' refers to H_2O, and the liquid to which I gesture is not H_2O. Now, suppose Twin Earth contains my molecular twin. This twin has grown up on Twin Earth, and experienced the same events as I have, including getting thrown into a wormhole to Earth. When he points at water and utters, 'That's water', he is also uttering a falsehood, because the substance to which he gestures is not XYZ. In other words, the extension of my use of 'water' is H_2O, while the extension of my twin's use is XYZ. Given that Fregeans hold that Sense determines reference, and that my twin and I are in the same psychological state, it follows that we must be grasping different Senses.[8] Therefore, there seems to be at least one way in which a simulation could provide evidence against a Fregean theory: if two agents have the same internal organization yet refer to different substances, they constitute a counterexample to the Fregean account.

The Fregean can effectively reply by noting the source of evidence is misplaced: the work is being done by the modal intuitions contained in the externalist thought experiment, not the simulation. What is required, as in the ideational and referential cases, is that the *simulation* exhibits a relevant counterexample to (a), but it is difficult to imagine how a simulation could possibly deliver such externalist intuitions independently of it being interpreted as such. Once again, adopting assumptions inconsistent with the theory being tested comes prior to interpreting the simulation.

The second case (b) fares no better. If we could show that Fregean theories were committed to the claim that two agents grasping the same meaning must instantiate the same psychological state, and furthermore that simulations instantiate agents grasping the same meaning while differing in psychological states, then the simulations would provide evidence against Fregean theories. However, first, there is no prima facie reason the set of psychological states that grasp a given Sense could not be disjunctive, so a Fregean could deny the first premise. Second, even if they were committed to the first premise, to show that simulated agents are grasping the same meaning we would need to interpret them as such, and this would require that we deploy a theory of meaning—in particular, one that is inconsistent with the Fregean version being challenged. Once again, the argument would beg the question.

Since on the Fregean account shared psychological states are neither necessary nor sufficient for grasping the same Sense, Fregean theories are not committed to *any* thesis about the mechanisms by which agents grasp Senses—they can simply assert that in order to solve such puzzles as substitution in opaque contexts there must be states constitutive of grasping for any creatures that understand a language. Therefore, as per the original strategy, Fregeans can claim that a simulation is an account of the causal mechanisms underlying linguistic behavior while denying that the simulation says anything about the viability of their theory of meaning. In other words, reason (i), above, is irrelevant because Senses do not play a causal role in the mechanisms of linguistic behavior; (ii) is false because there *is* a grasping relation instantiated by the agents; and (iii) is misconstrued, as the two notions of successful communication—as grasping the same Sense and as coordinating behaviour—can be simultaneously true.

Now, to the simulation proponent this may continue to seem like a *reductio*: if meanings are mind-independent Senses, and meanings are involved in communication then claiming that a simulation of linguistic behavior need not represent Senses is absurd. But this ignores the point: the Fregean (or at least the variety I'm considering here) disagrees with the assumption that Senses have *any* causal role to play in linguistic behavior. Instead, the psychological states involved in grasping Senses are what play a causal role. So, while the simulations may be relevant to constructing an adequate account of how agents grasp Senses, they are not relevant to the theory of meaning itself. In this way the Fregean theory remains consistent with the simulations.

Furthermore, a familiar general conclusion can be drawn from the preceding discussion: Assessing simulation results as relevant for the Fregean theory involves appealing to evidence independent of and prior to the simulation, either through appealing to intuitions based on our firsthand linguistic knowledge (as in Twin Earth thought experiments), or by presupposing how to interpret the meaningfulness of the simulation behavior (as in the discussion of case (b)). Once again, the simulations themselves do not seem have any purchase on the theory of meaning under consideration.

If this is correct, then we've arrived at the following general conclusions. First, against the claim that GSK's simulations provide support for 'use' theories of meaning, closer inspection reveals that the results are in fact consistent with all of the theories of meaning surveyed by the authors, and hence do not favor any one account. Second, theories of meaning (as represented by the ideational, referential, and Fregean accounts) are not the kinds of theories for which simulation results can provide evidence. For each of the theories considered, taking simulation results to be relevant requires that we first interpret those results, but doings so involves begging the question against alternative interpretations based on other theories of meaning, or otherwise 'importing' the phenomenon of interest from a prior source—our own language and intuitions.

If theories of meaning *are* immune to testing through computational simulation, then one might think there are consequences for linguistic naturalism. I conclude by articulating this worry, and arguing that if there are any such consequences, they are minimal.

4. Conclusions

Roughly speaking, linguistic naturalism is the view that any adequate theory of language will appeal only to entities, events, properties, or kinds dealt with by the natural sciences; no 'supernatural' or 'non-natural' properties are necessary. In addition to this metaphysical requirement, naturalism is often defined as having a methodological component: theory testing and construction should make use only of naturalistic methods, i.e., the methods of natural science.

On a standard reading, Fregean theories are non-natural since they fail (at least) the metaphysical requirement by positing immaterial Senses, so let us bracket them for the remainder of the paper. In contrast, ideational and referential (causal) theories are typically understood as metaphysically naturalistic. However, given that computational simulations constitute a legitimate naturalistic *method*, does the claim that these theories of meaning resist investigation through simulations suggest that, regardless of their metaphysical status, such theories are *methodologically* non-natural? It seems we can easily resist an affirmative answer to this question.

Computational simulations constitute only one tool in the very large toolbox of naturalistic methodology, and in the present context its elimination leaves plenty of other respectable methods upon which to found a naturalistic theory of meaning. For example, the reports of competent speakers of a language constitute respectable empirical data in linguistics, and this same data is often relevant to the evaluation of theories of meaning. Furthermore, many philosophers consider the thought experiments utilized in debates over theories of meaning—such as Putnam's externalist intuition pump—are themselves a naturalistic method. Therefore, a proponent of metaphysically and methodologically naturalistic theories of meaning need not discard the methodological component simply because computational simulations do not do the trick.

To sum: attempts to use computational simulations of language emergence as sources of empirical evidence for evaluating theories of meaning face important challenges. The first type concerns the epistemological role of simulations: what conditions must be met in order that inferences from simulation results to target system are justified? I've suggested that the simulation proponent must find a way to situate simulation models between mere demonstrations of possibility and robustly relevant simulations, or otherwise articulate an epistemological role that justifies the claim that simulations can provide 'support'. While this task may be difficult, it is not obviously impossible (and furthermore, there are other interesting epistemological roles for computational simulations that deserve investigation in their own right). The second type of difficulty concerns the theories to be tested by simulation studies: I've argued that GSK's results are consistent with all of the theories of meaning surveyed by the authors, and hence do not single out 'use' theories for support. More fundamentally, theories of meaning (as represented by the ideational, referential, and Fregean accounts) do not appear to be the kinds of theories for which simulation results can serve as evidence. Consolation may be taken from the fact that this conclusion has little impact on the claim that (at least some) theories of meaning are methodologically naturalistic.

Notes

[1] Variations on these putative virtues are also found in Boden (1977), Cangelosi and Parisi (2002), and Webb (2001).

[2] The purposes of this paper do not require that we go into much detail regarding these competing theories of meaning. This means that I will not bring up many of the multitude of interesting issues that surround them. This also means I must issue a caveat: while I believe the arguments presented below work with these types of theory generally construed, it may nonetheless be the case that specific elaborations on a theory may present exceptions. I leave the detection and handling of any exceptions as a future task; the present goal is to explore the larger issue of the relationship between simulation and theories of meaning.

[3] Some of the simulations may not fit precisely into this characterization, but it nonetheless captures the spirit of those simulations. For example, Steels and Vogt's experiments reported in their (1997) take shared perceptual representations as the criterion for successful communication. So, for example, a speaker X successfully communicates with a listener Y

when, as a result of hearing X's utterance, Y tokens internally the same set of perceptual features that X tokened upon viewing object O and which caused X to make that utterance. However, the overall perspective governing the experiments is (late) Wittgensteinian insofar as their emphasis is on X and Y successfully picking out the same object from a set of possible objects, a task that does not obviously require shared internal representations (and which at least in some cases appears to be achieved by the agents in his experiment prior to their having shared internal representations).

[4] Because of contingencies involving (i) initial conditions, (ii) which neighbors are selected to train an agent over the course of the simulation, (iii) the movement of food and predators, and (iv) which subsets of a neighbor's input-output profile are used to train an agent, different 'dialects' may develop: a single environment can contain multiple spatially-contiguous groups, each of which utilizes its own lexicon.

[5] The objection can also be adapted to Kripke-style causal chains. Suppose that the reference of a term is established by a 'baptismal' event, where an agent bestows a name upon a type of object. This name is then passed along to other agents, forming a causal chain from any current use of a term back to the original object. GSK's simulations can be interpreted in a way consistent with this causal-chain account. Since the agents are randomly initialized, the community begins with multiple distinct naming events for the same type of object. This sets up a population of competing causal chains of reference, and through processes of selection during the course of the simulation, one chain eventually emerges as dominant. So the simulation is compatible with basic causal-chain as well as causal co-variation theories.

[6] At this point it should be clear that the argument of this paper makes extensive use of thought experiments. This prompted an interesting question on the part of an anonymous reviewer, which, if I interpreted it correctly, goes something like: computational simulations are a form of thought experiment; and you argue against their utility; but how can you use thought experiments to argue that thought experiments fail to be useful? On the one hand, this question brings up issues beyond the scope of this paper: are simulations thought experiments, are they elucidations and implementations of thought experiments, or are they something else? If they are thought experiments, are they of the same type as other thought experiments? These are very general and important questions about the nature of computational simulations and thought experiments. On the other hand, I believe that the thesis of this paper requires only the intuitive and unproblematic distinction between (i) thought experiments that probe a person's semantic intuitions directly, and (ii) the process of observing and interpreting simulation behavior qua linguistic behavior.

[7] Obviously, for a simulation to be relevant to anything it is going to have to be interpreted. The objection is not that simulation results are undermined merely by the need to interpret them. The objection is that, in the case of theories of meaning, the simulation results are undermined because interpreting them as relevant to the debate requires that we assume without justification that the simulation exhibits phenomena problematic for the theories we are trying to evaluate.

[8] Alternatively, a Fregean might conclude that my twin and I grasp the same Sense (by virtue of instantiating the same psychological state), but a single Sense can determine multiple extensions (depending on other factors). Putnam notes that this interpretation leads to the possibility of different words having the same meaning yet having different extensions, which is dismissed as "highly counterintuitive" (Putnam, 1973, p. 710, n. 2). Here I follow suit.

References

Boden, M. (1977). *Artificial intelligence and natural man*. New York: Basic Books.

Cangelosi, A., & Parisi, D. (2002). Computer simulation: A new scientific approach to the study of language evolution. In A. Cangelosi & D. Parisi (Eds.), *Simulating the evolution of language* (pp. 3–28). London: Springer.

Dyer, M. G. (1995). Toward synthesizing neural networks that exhibit cooperative intellegent behavior: Some open issues in artificial life. In C. G. Langton (Ed.), *Artificial life: An overview* (pp. 111–134). Cambridge: MIT Press.

Fodor, J. A. (1990). A theory of content I: The problem. In *A Theory of Content and Other Essays* (pp. 51–87). Cambridge, MA: MIT Press.

Grimm, P., St. Denis, P., & Kokalis, T. (2004). Information and meaning: Use-based models in arrays of neural nets. *Minds and Machines*, *14*, 43–66.

Hutchins, E., & Hazlehurst, B. (1991). Learning in the cultural process. In C. G. Langton, C. Taylor, J. D. Farmer, & S. Rasmussen (Eds.), *Artificial Life II* (pp. 689–708). Redwood City: Addision-Wesley.

Norton, S. D., & Suppe, F. (2001). Why atmospheric modeling is good science. In C. A. Miller & P. N. Edwards (Eds.), *Changing the atmosphere: Expert knowledge and environmental governance* (pp. 67–105). Cambridge, MA: MIT Press.

Putnam, H. (1973). Meaning and reference. *The Journal of Philosophy*, *70*, 699–711.

Steels, L., & Belpaeme, T. (2005). Coordinating perceptually grounded categories through language: A case study for colour. *Behavioral and Brain Sciences*, *28*, 269–529.

Steels, L., & Vogt, P. (1997). Grounding adaptive language games in robotic agents. In P. Husbands & I. Harvey (Eds.), *Fourth European Conference on Artificial Life* (pp. 474–482). Cambridge, MA: MIT Press.

Webb, B. (2001). Can robots make good models of biological behaviour? *Behavioral and Brain Sciences*, *24*, 1033–1050.