Behavioral Foundations for Expression Meaning

Megan Henricks Stotts

This is a post-peer-review, pre-copyedit version of an article published in *Topoi*. The final authenticated version is available online at: http://dx.doi.org/10.1007/s11245-019-09675-0.

Abstract:
According to a well-established tradition in the philosophy of language, we can understand what makes an arbitrary sound, gesture, or marking into a meaningful linguistic expression only by appealing to mental states, such as beliefs and intentions. In this paper, I explore the contrasting possibility of understanding the meaningfulness of linguistic expressions just in terms of observable linguistic behavior. Specifically, I explore the view that a type of sound (or other item) becomes a meaningful linguistic expression within a group in virtue of the production of that type of item becoming that group's widespread, copied way of getting others to involve an object or relation in their activity. After discussing a preliminary version of the view, I develop it in response to some key concerns about whether it really does, as a matter of fact, eschew mental states, and about its adequacy as an account of linguistic meaning.

1. Introduction

Some sounds that the human mouth can produce have the interesting feature of being publicly meaningful linguistic expressions. This feature, which I will call *expression meaning*, is also shared by some gestures and markings. A *foundational theory of meaning* seeks to identify the facts in virtue of which sounds and other items have this kind of public, group-level meaningfulness.[1] According to well-established philosophical tradition, the correct foundational theory of meaning must appeal to language users' mental states: that is, linguistic expressions are thought to be meaningful partly or wholly in virtue of people's intentions, beliefs, or other mental states or processes.[2]

In contrast to this well-established tradition, I'd like to explore the possibility of a different kind of foundational theory of meaning. Consider human linguistic behavior, observed from the outside. We produce a wide variety of sounds, gestures, and markings, with identifiable patterns:

[1] For more on the notion of a foundational theory of meaning and its differences from other semantic projects, see Kaplan (1989, pp. 573–574), Stalnaker (2003, pp. 166–167), Williams (2007, p. 361), García-Carpintero (2012a, p. 397), Burgess and Sherman (2014, pp. 1–2), Speaks (2015), and Simchen (2017, p. 175). Many of these authors also (or exclusively) use the term 'metasemantics' for the project of giving a foundational theory of meaning. I prefer the term 'foundational theory of meaning' because metasemantics can also be thought of as a broader enterprise that includes other work that is in some sense prior to semantic theory, such as, for instance, work on the metaphysics of propositions (*cf.* Burgess and Sherman 2014).

[2] For example, Stephen Schiffer (1972) and Jonathan Bennett (1976) build on work by Paul Grice (1989) to argue for foundational theories of meaning that appeal to speakers' intentions. Other foundational theories of meaning that appeal to mental states have been suggested by David Lewis (1975), Brian Loar (1976), Donald Davidson (1973) (at least as Manuel García-Carpintero (2012a) interprets him (pp. 403–404)), Michael Dummett (1996), Paul Horwich (1998, 2004), Wayne Davis (2003, 2005), John Hawthorne (2007), García-Carpintero (2012b), and Josh Armstrong (2016b).

many of the same items appear and reappear in different combinations. As language users, we know that there are mental states—desires, intentions, beliefs—behind much of that behavior. But I'd like to consider the possibility of a foundational theory of meaning that makes no appeal to the mental states behind the scenes of our linguistic behavior, which I'll call the Behavioral Theory of Meaning (BTM). As its name suggests, the BTM will aim to explain the meaningfulness of linguistic expressions in terms of just observable linguistic behavior and some other closely connected non-mental phenomena. It's important to note that taking a behavioral approach to expression meaning would not amount to taking a behavior*ist* approach to the philosophy of language more broadly. Even if we come to see expression meaning as entirely behavioral, it is still possible to see the broader phenomenon of linguistic communication (which lies outside of this paper's scope) as necessarily involving speakers' and hearers' mental states.[3]

The overall aim of the present paper is to introduce the BTM and then show how it can respond to some key concerns about whether it is, as a matter of fact, entirely behavioral, and about its adequacy as a foundational theory of meaning. We'll begin, in Section 2, by discussing a preliminary version of the BTM, and we'll consider two concerns about whether it actually is entirely behavioral in Section 3. Then, in Sections 4 and 5, we'll see that the preliminary version of the BTM over-generates expression meaning in several ways. Discussion of these over-generation problems will lead us to modify the BTM, culminating in a revised version of the theory at the end of Section 5. In Section 6, we'll discuss a concern that the BTM may be unable to accommodate the phenomenon of semantic deference.

[3] For more on this kind of approach to the broader phenomenon of linguistic communication, see Stotts (forthcoming).

2. A Preliminary Version of the Theory

In this section, we'll introduce a simple, preliminary version of the BTM. Many elements of the preliminary view will then require elaboration, with particular attention to the question of whether the theory really is entirely behavioral.

But first, a note about the BTM's scope of application is in order. As presented in this paper, the BTM is designed to apply only to morphemes that have relations (including properties) or objects as their meanings. I use 'object' and 'relation' in their widest construals and with minimal theoretical commitments. Any "thing" counts as an object—a tree, Iceland, the number 4, a pain, a thought, or a crowd. Some objects are physical; some are abstract; some are mental.[4] Any property of objects, or relationship among objects, counts as a relation—being green, being a member of a certain species, being the favorite food of some other object, being the square root of another object, *etc.* As discussed in this paper, the BTM will not apply to expressions such as, for instance, the English suffixes '-s' and '-ing,' which do not have objects or relations as their meanings. Instead, we'll be concerned just with proper names, common nouns, and adjectives. I plan to explore ways of expanding the BTM to apply to morphemes of all kinds, and to sentences, in future work.

The basic idea behind the BTM is that expression meaning arises when the production of some arbitrary item (such as a sound) is copied within some group as a way of getting others to engage in activity involving some object or relation. But the preceding is rather vague. Here is a first pass at spelling out the details:

[4] Thus, the BTM's aim of giving an entirely behavioral, non-mental story about the facts in virtue of which linguistic expressions are meaningful does not preclude it from capturing the fact that some words have mental objects as their meanings.

(BTM 1) A type of item *I* means object or relation *O* within a group in virtue of both of the following conditions being satisfied:

(1) Within the group there is widespread, interconnected copying of producing *I* as a way of getting others to involve *O* in their activity.
(2) There is at least one other way of getting others to involve *O* in their activity that is at least approximately as conducive to that effect and at least approximately as accessible to the group as producing *I*, independent of the dominance producing *I* has gained due to being copied.

For example, consider the meaningfulness of the sound >green< within English-speaking groups.[5] In accordance with condition (1) of (BTM 1), people copy each other in producing that sound as a way of getting others to engage in various kinds of activity involving the property of giving off light with a wavelength close to 500 nanometers—for instance, some productions of >green< may aim at getting someone to make something have that property (perhaps by applying paint), whereas others may aim at getting someone to avoid something that has that property (perhaps by hiding from someone wearing green). And in accordance with condition (2), before >green< became dominant, there were many other sounds that could just as easily have come to have that meaning instead, due to the fact that >green< was initially no likelier than its many alternatives to succeed in getting people to involve that property in activity.

Several terms that appear in (BTM 1) require discussion, starting with the notion of a "type of item." The term 'item' is intended to have very broad application—a sound, a gesture, a marking, or even some other kind of physical object can become a meaningful linguistic expression. Crucially, it's not just a *particular* item that becomes a meaningful expression—rather, a *type* of item must become meaningful within a group in order for the linguistic expression to be producible by multiple individuals. This means that we need to say something about what makes two items tokens of the same type. In my view, the most promising proposal is to sort item tokens into types on the basis of their causal histories.[6] First, consider sounds: although there are noticeable differences in the

[5] I use reversed angle brackets (> <) to designate the type of *sound* that corresponds to the spelling within the brackets. The issue of how to individuate sound types will receive discussion below.
[6] Here I am inspired by Ruth Millikan (2005), who divides words into types by their copying lineages (pp. 33–34, 60–62).

pronunciation of >bagel< in different regions of the United States, all of these pronunciations count as tokens of the same sound because of their shared origins—they come from branches of a single causal tree, linked together by people copying (however imprecisely) others' pronunciation over time. Relying on shared histories in this way saves the BTM from the vexing question of exactly how acoustically similar two noises must be to count as tokens of the same sound. Gestures, markings, and other items can be divided into types by their histories as well.

It's important to note that two sounds (or other items) with one *component* that has a shared history do not satisfy the criterion just introduced. For instance, >candle< and >candid< can both be traced back to the Latin >candere<, but presumably not as two branches of a single, unified tree. Rather, >candere< was just one influence on >candid< and on >candle<, and each sound's ending has its own separate source. So >candid< and >candle< are two different sounds with one shared component, rather than being different pronunciations of a single sound.[7]

The notion of an object or relation being "involved in someone's activity" also requires elaboration. 'Activity' is a technical term here, and its scope includes mere mental activity (*e.g.*, believing, thinking, desiring), as well as physical movement. (Behavior, then, is activity that is directly observable by others.) At the level of sentences, expressions get people to involve objects and relations in many different *kinds* of activity, from believing that an object stands in some relation, to making an object stand in a relation, to preventing an object from standing in a relation.

The notion of an object or relation being *involved* in someone's activity is intentionally quite permissive, because what it is for an object or relation to be involved in someone's activity varies with the nature of the activity, and with the nature of the object or relation. For instance, an object is

[7] I should note that if, as a matter of fact, >candle< and >candid< did come from >candere< by means of a simple shift in pronunciation (as with the variant pronunciations of >bagel<) rather than through the influence of two other chains of copying, then >candle< and >candid< *would* be the same sound, pronounced differently. We might operate under the belief that they are two different sounds, but we would just be wrong about that feature of our language.

involved in my activity of believing if it is part of the content of my belief, and this applies just as well to abstract and mental objects as to physical ones.[8] But an object is involved in my activity of sitting only if my behavior causally impacts it in a certain way, which is certainly not something that could be said of an abstract object. I won't say more about the metaphysical question of what it is for an object or relation to be involved in someone's activity, but I hope the preceding is enough to make that notion sufficiently clear. It's also important to emphasize that the rather bare notion of getting someone to involve an object or relation in activity is not intended to be a gloss for successful communication. Rather, it is just one element of the much richer phenomenon of communication, which lies beyond the scope of this paper.

The allowance above that the activity that the production of some item aims to cause can be mental activity may seem to make it impossible for (BTM 1) to actually be entirely behavioral. However, we can see that this is not the case if we recognize that the audience's mental activity constitutes only their *understanding* of the utterance and their further responses to it; it does not play a role in the BTM's picture of what constitutes expression meaning itself. The BTM aims to give a behavioral account of what makes linguistic expressions meaningful, not of the comprehension process or other further effects on hearers. Because the *speaker's* side is what gives rise to expression meaning, *it* is what the BTM aims to portray as entirely behavioral.[9] So, according to the BTM, expression meaning arises when speakers produce some item *as a way of* bringing about some effect on hearers' activity; whether that effect actually occurs is irrelevant to the metaphysical determination of the item's meaning. Near the end of Section 3, we'll be able to say something a bit more precise about this feature of the BTM.

[8] I should note that here I am assuming a broadly externalist notion of mental content according to which actual objects and relations in the world are included in the content of one's mental states.

[9] This means that the BTM is a *productivist* (rather than *interpretationist*) foundational theory of meaning, in Ori Simchen's (2017) terminology.

Next, we'll discuss conditions (1) and (2) of (BTM 1), starting with condition (2) for reasons of expository ease. Condition (2) reads as follows: "There is at least one other way of getting others to involve *O* in their activity that is at least approximately as conducive to that effect and at least approximately as accessible to the group as producing *I*, independent of the dominance producing *I* has gained due to being copied." This statement captures the fact that expression meaning is *arbitrary* in a very particular sense. The human mouth can produce a wide range of sounds, and none of them are inherently meaningful. For example, we could have used thousands of sounds aside from >green< to talk about that particular color, but somehow >green< is the one that took off. This is a key feature of expression meaning that a foundational theory of meaning must capture, because it is a central difference between expression meaning and the significance of a mere natural sign or symptom. Spots of a certain kind just do mean measles; there are no alternative kinds of spots that could have indicated measles instead (at least, not within the realm of physical possibility). But the meaning of a sound such as >green< could have attached to countless other sounds instead (Grice 1989, pp. 213–214).

It is crucial to require, as condition (2) does, that the alternatives to the production of an item that has become meaningful be at least approximately as conducive to the relevant effect and at least approximately as accessible as the item that is in use. An inferior alternative does not make a practice arbitrary—for instance, there are alternatives to watering that are less conducive to keeping plants alive, such as pouring milk on them, but that does not make watering plants an arbitrary practice. On the other hand, the practice of putting gasoline in one's car may have some equally good alternatives—perhaps there are other physically possible substances that would be just as conducive to making cars run—but those alternatives do not make putting gasoline in one's car arbitrary because they are so much less accessible to us than using gasoline is.[10]

[10] See Stotts (2017) for more on these notions of accessibility and arbitrariness.

Moreover, condition (2) requires that the alternatives be equally conducive to the relevant effect and equally accessible only "independent of the dominance producing *I* has gained due to being copied." We'll say more about copying in a moment. This qualification on condition (2) serves to recognize the fact that once an item's meaningfulness becomes well established in a group, producing that item is often by far the best and most accessible way of getting people involved with the relevant object or relation (*cf.* Burge 1975, p. 254; Millikan 2005, p. 57). This does not mean that linguistic expressions are not arbitrary; it just means that the arbitrariness at issue is arbitrariness independent of the dominance they gain when their use becomes widespread. So, what makes the production of >green< arbitrary in the right way is that *before* that sound took off, there were equally accessible alternatives that were equally conducive to the relevant effect.

Interestingly, the arbitrariness of linguistic expressions is really an *obstacle* to their meaningfulness: because there are equally good, equally accessible alternatives, it is difficult for any one of those alternatives to take off. (BTM 1)'s condition (1) aims to explain how that obstacle is overcome. As we can see in condition (1) (which reads: "Within the group there is widespread, interconnected copying of producing *I* as a way of getting others to involve *O* in their activity"), the notion of *copying* is the key to the explanation. Here I'm using a notion of copying adapted from Ruth Garrett Millikan's (2005) notion of reproduction: a current instance of behavior is a copy of a past one if it shares some features with the past behavior, and the past behavior caused the present behavior to have those shared features (p. 3).[11] As Millikan (2005) emphasizes, copying can be completely

[11] Millikan's (2005) definition includes a counterfactual component: "For one thing to be 'reproduced' from another, all that is required is that there be a mechanism that produces the second on the model of the first, such that *had the first been different in specifiable respects, that would have caused the second to differ accordingly*" (p. 31, my italics). Millikan's counterfactual component is too strong. I might, for instance, copy someone's dance moves without it being the case that I would still have copied her if she had engaged in *different* dance moves. All that's needed is an actual causal connection between the way the original dancer moves, and the way I move. I should also note that Millikan (2005) uses the term 'direct copying' for just one particular kind of reproduction, in which one individual directly observes and imitates another (pp. 4–5). My notion of copying is broader than this—it encompasses any situation in which the right causal connection obtains between two instances of behavior. However, I do emphasize the role of direct copying more than Millikan does, which makes the terminological shift apt.

unintentional and unconscious (pp. 5–6). When I sit up straight at a restaurant with my knees tucked under the table, that behavior is partly caused by past instances of others sitting in that way in restaurants, whether I realize this or not. Copying, then, has to do with causal connections among behavior, and not with anyone's mental states.

The requirement within condition (1) that the copying be *widespread* comes from the recognition that in order for expression meaning to arise, it is surely not sufficient that the use of some item be copied once or twice, but on the other hand, there is no exact threshold for how many times it must be copied. To make the term 'widespread' more precise, we can say that a practice is widespread in a group provided that most members of the group have encountered that practice, even if they have not engaged in it themselves. It is also important to require that this widespread copying be causally *interconnected*, in order to rule out cases in which, for unrelated reasons, many group members end up engaging in the same kind of linguistic behavior and then are each copied once. That would be widespread copying, but it would just be a coincidence within the larger group.[12]

A final feature of (BTM 1) worth mentioning is that it (along with the revised versions of the BTM below) can be seen as an account of expression meaning in terms of *social conventions*. I have not made this claim explicit in the BTM because it relies on a controversial approach to conventions coming from Millikan (2005), according to which conventions are entirely behavioral. This goes against the dominant view that conventions necessarily involve mental states (*e.g.*, Lewis 1969/2002, 1975; Gilbert 1989; Miller 2001; Marmor 2009). My aim in the present paper is just to explore the plausibility of the BTM; the question of whether the BTM implies that expression meaning is conventional is separable.

[12] I am grateful to Monique Wonderly for this point.

3. Effect-Sensitive Behavior

In the previous section, we gave some attention to the issue of whether (BTM 1) really is as behavioral as it aims to be, when discussing the notions of copying and activity. But a subtler threat to (BTM 1)'s behavioral aspirations lurks in the notion of producing some item *as a way of* getting others to involve some object or relation in their activity.[13] If I say, for instance, that I left the peanuts out of a batch of cookies as a way of protecting an allergic friend, it certainly sounds as if I'm describing myself as having acted with a certain *intention*. If the 'as a way of' locution must ultimately be understood in terms of intentions, the BTM will be unable to live up to its name.

To allay the worry that the 'as a way of' locution smuggles intentions into the BTM, we'll isolate a particular kind of behavior: effect-sensitive behavior. Broadly speaking, effect-sensitive behavior is behavior that observably tracks one of its possible effects. We'll identify exactly what it means for behavior to be sensitive to an effect, and then show that this does not imply that the effect must be *intended*.

Imagine someone walking down a crowded sidewalk. She sometimes weaves as she's walking, and she is much likelier to weave when weaving is conducive to avoiding collisions with other pedestrians. She may not always weave when it would allow her to avoid bumping into others, but nonetheless she's *likelier* to weave when it's conducive to avoiding collisions. Now imagine that the sidewalk becomes so crowded that weaving is much less likely to allow her to avoid collisions. Suppose that then her weaving behavior lessens noticeably, although she may still weave occasionally, perhaps when there is broken glass in her path. If we watched her over time, we could observe that her weaving behavior is *tracking* the effect of avoiding collisions. The weaving behavior always has plenty of other effects (*e.g.*, annoying someone walking behind her, or crushing an ant), but the effect of avoiding collisions seems to explain the behavior in a way that the others do not.

[13] My use of the 'as a way of' locution is inspired by Israel, Perry, and Tutiya (1993).

Here is a general definition of this kind of effect-sensitive behavior:

An individual's behavior *B* is sensitive to an effect *E* if both of the following conditions are satisfied:

(1) Increases in the likelihood that *B* will cause *E* correlate with increases in the likelihood that *B* will continue or intensify.
(2) Decreases in the likelihood that *B* will cause *E* correlate with increases in the likelihood that *B* will lessen or cease.

According to this definition, behavior must persist over time in order for sensitivity to an effect to manifest. Now we'll discuss a second kind of effect-sensitive behavior that can accommodate behavior that is short-lived on any particular occasion, but repeated on multiple occasions. Think of a professional baseball player's behavior of swinging the bat when a pitcher throws a ball toward him. If we were to observe him at bat on many different occasions, we would see that the occasions on which swinging the bat has a higher chance of resulting in hitting the ball (*i.e.*, when the pitcher throws the ball at the right height and angle) are also occasions on which the batter is likelier to swing. Moreover, occasions on which swinging the bat has a relatively low chance of hitting the ball are occasions on which the batter is less likely to swing. Of course, sometimes the batter will get a strike: he will sometimes swing even when it isn't likely to result in hitting the ball, or refrain from swinging even when swinging is likely to result in hitting the ball. But overall, his swinging behavior tracks the effect of hitting the ball.

Now for the official definition of this second kind of effect-sensitive behavior:

An individual's behavior *B* is sensitive to an effect *E* if all of the following conditions are satisfied:

(1) *B* occurs in a context in which it is relatively likely to cause *E*.
(2) In the past, occasions with relatively high likelihood that behavior of the same kind as *B* would cause *E* have been occasions on which the individual was also relatively likely to engage in that kind of behavior.
(3) In the past, occasions with relatively low likelihood that behavior of the same kind as *B* would cause *E* have been occasions on which the individual was also relatively unlikely to engage in that kind of behavior.

Here, 'relative likelihood' means likelihood relative to most of the situations in which the individual has previously found herself. Conditions (2) and (3) are analogous to the conditions for the first kind of effect-sensitive behavior, but condition (1) is new. We need condition (1) because without it, our

baseball player's swinging behavior when he is warming up with the bat with no pitcher in sight could count as sensitive to the effect of hitting a ball. More generally, when behavior occurs on a new occasion, we don't have any reason to think that it shares past similar behavior's sensitivity to some effect unless that effect is relatively likely in the current context.

This second kind of effect-sensitive behavior is more applicable to linguistic behavior than the first, because our linguistic behavior involves producing the same expressions on different occasions. Imagine that a professor in the United States produces the sound >Seattle< in the presence of some of her students. In that context, producing >Seattle< is likely to get people to involve a particular U.S. city in some kind of activity. Condition (1) is satisfied. It's easy to imagine that in the past, the professor has been likelier to produce >Seattle< in contexts in which it was likely to get people to involve that city in activity—that is, in contexts that included humans who were likely to have past exposure to that sound. And she has been less likely to produce it in contexts in which it was unlikely to have that effect, such as when she was among people who have never visited North America and are unlikely to have encountered anyone who has. Conditions (2) and (3), then, are satisfied as well.

So, behavior can manifest sensitivity to some effect either over the course of one extended episode, or over the course of multiple episodes of behavior of the same kind, as in linguistic behavior. We can describe behavior as *more* or *less* sensitive to some effect, depending on how much its likelihood (or intensity) increases when the likelihood of the effect increases and how much its likelihood decreases (or how much the behavior lessens) when the likelihood of the effect decreases.

Carving out this kind of behavior is all well and good, but what we really need to know is whether effect-sensitive behavior implies the existence of intentions. Intuitively, in the initial cases we discussed, it seems that intentions are not necessarily present. The woman walking down the street might not even realize that she is weaving, let alone intend to do it, and the batter's habits might be so deeply ingrained that he doesn't think about what he's doing. But it's easy to push back on those

intuitions, insisting that even if these agents are not *conscious* of having an intention, they might nonetheless have one.

For better evidence that effect-sensitive behavior does not imply a corresponding intention, let's consider a case in which a person's behavior is sensitive to an effect with a clear absence of an intention to bring about that effect. Think of a person who makes an offer on some real estate. When others make offers on the land, she increases her own, and if interest in the property begins to lag, she decreases what she is willing to pay. Her behavior is sensitive to the effect of acquiring the property. But imagine that she actually intends *not* to acquire the property; she intends to prevent another person from acquiring it, and she plans to drop her offer as soon as that person gives up. Her behavior is sensitive to the effect of acquiring the property, but she intends *not* to acquire the property. Thus, a person's behavior can be sensitive to some effect without her intending to bring about that effect.

And in fact, we can turn to another kind of case to see that not only does effect-sensitive behavior not require an intention to bring about the effect to which the behavior is sensitive, but it does not require intentions at all. Here we can look to organisms that are incapable of robust intentions but can nonetheless engage in effect-sensitive behavior. Think of a dog whining as long as that behavior conduces to the effect of getting attention, and ceasing whining when whining is no longer conducive to getting attention. Or, perhaps more strikingly, think of ants' food-seeking behavior. Ants are likelier to move in a particular direction if it is conducive to getting food, and less likely if it is not. If effect-sensitive behavior required intentions, only beings that can form intentions would be able to exhibit it.

None of this requires us to deny that, at least for humans, intentions and other mental states often accompany effect-sensitive behavior. When I leave the peanuts out of some baked goods to protect an allergic friend, my effect-sensitive behavior is guided by an intention. But even then, the notion of effect-sensitivity allows us to isolate a purely behavioral layer of my activity: if you observe

me over time, you will see that I am likelier to leave the peanuts out when my allergic friend is likely to eat the baked goods, and less likely to leave them out when he is not likely to eat them. Intentions and effect-sensitive behavior come apart—both in the sense that each can exist without the other, and in the sense that even when they co-occur, they are separable. So, relying on the notion of effect-sensitive behavior, we can see that the 'as a way of' locution does not smuggle intentions into (BTM 1).

But, there is a second possible challenge to the BTM's ability to portray expression meaning as entirely behavioral that might seem to accompany our notion of effect-sensitive behavior. Our definitions of effect-sensitive behavior make reference to the *likelihood* of various effects in various contexts. For the BTM, the effects whose likelihood matters are effects on hearers' activity (such as getting hearers to involve a certain property in activity when they hear >green<), which we've already noted may include mental activity (such as believing that something is green). So, although the definitions of effect-sensitive behavior do not appeal to the actual occurrence of mental phenomena, they do appeal (once applied to linguistic behavior) to the *likelihood* of some mental phenomena. If the phenomenon of some mental activity being likely is itself a mental phenomenon, then effect-sensitive behavior will have built an implicit appeal to mental phenomena into the BTM.

This concern may seem especially pressing when we consider that for some linguistic expressions in some groups, mental activity is the *only* kind of activity in which members can involve the relevant object or relation. Think, for example, of a group of English-speakers who use >Mount Kilimanjaro< as a way of getting others to involve that particular mountain in activity, but none of them ever visit that mountain nor do anything to physically impact it. Mental activity will then be the *only* kind of activity whose likelihood will make it the case that the behavior of speakers in that group satisfies the second definition of effect-sensitive behavior.

However, the phenomenon of some mental activity being likely in a given situation is not itself a mental phenomenon. That I think about the property of greenness is something that goes on inside my head; that it is *likely* that I would think about that property if you were to utter >green< is *not* something that goes on inside my head. Rather, the likelihood of a mental phenomenon resulting from a possible kind of behavior in a given situation can be cashed out just in terms of the presence of humans who have previously encountered certain objects and utterances. This, in fact, was how we handled the matter in our discussion of >Seattle< above: we cashed out the likelihood of the effect of getting others to involve Seattle in their activity in terms of the presence of North American humans who had encountered >Seattle< before. It is the presence of those people and the history of what sounds they have encountered that makes it the case that certain mental phenomena are likely. These sorts of facts about who is present and what they have encountered in the past are not mental phenomena, and thus the likelihood of mental phenomena occurring is not itself a mental phenomenon. The point applies to the >Mount Kilimanjaro< example, too: what makes it likely that uttering >Mount Kilimanjaro< will produce the effect of getting others to involve Mount Kilimanjaro in activity is the presence of human beings who have encountered other copies of that sound in the past and who have, perhaps, encountered pictures or maps of Mount Kilimanjaro. So, (BTM 1) still appears to have its behavioral credentials intact.

The preceding discussion allows us to be a bit more precise about the point from Section 2 about the BTM's focus on just the speaker, and not on the actual effect on the hearer (if any). We're now in a position to say that according to the BTM, what gives rise to expression meaning is not anything hearers do (whether mental or otherwise) in response to utterances, but rather it is just the fact that speakers within a given group copy each other in producing a certain item in contexts in which doing so is likely to have a certain kind of effect, provided that speakers are less likely to produce that item in contexts in which that kind of effect is less likely. The effect itself may involve mental

phenomena in two different ways: on particular occasions, the kind of activity in which the hearer is to involve the object or relation may be mental activity (such as imagining an object or forming a belief about it), or the object or relation itself might be mental in nature (such as is the case for >pain< and >thought<). But the *likelihood* of that effect is all that matters for characterizing the speaker's behavior, and that likelihood is just a matter of non-mental features of the situation such as who is present and what meaningful linguistic items they have encountered in the past.

Before moving on to some problems with (BTM 1) in the next section, it's worthwhile to note that despite its non-Gricean emphasis on behavior, the BTM is influenced by Grice. In my view, Grice's (1989) most valuable insight about expression meaning was that the meaningfulness of linguistic expressions is closely connected to the sorts of effects that speakers aim to have on hearers when they utter those expressions. He develops this insight by claiming that expression meaning is constituted by speakers' communicative *intentions* (1989, p. 220). Instead of saying that linguistic expressions inherit their directedness toward the world from the directedness of the mental states that *guide* our use of those expressions, the BTM claims that linguistic expressions inherit their directedness from the public, observable directedness of our *behavior* toward certain possible effects on our audience. The BTM benefits from Grice's insight, but offers a behavioral spin.

Moreover, like Grice's theory, the BTM can be thought of as a kind of use theory of meaning, broadly speaking. Wittgenstein (1953/2001) suggested that meaning *is* use (§43), whereas Grice's theory of meaning and the BTM treat meaning as *determined* by use. This difference matters because the idea that meaning is just *determined* by use is perfectly compatible with the view that the meaning of an expression is (at least in some cases) "the object for which the word stands," which is an idea with which Wittgenstein (1953/2001) contrasted the idea that meaning *is* use (§1). Another way to describe the difference between Grice's theory and the BTM is that they portray meaning as determined by different *aspects* of use. For Grice, the mental aspect of use (*i.e.*, the intentions that

guide use) determines meaning. For the BTM, the observable, behavioral aspect of use determines meaning.[14]

4. The First Unrestricted Effect

In this section and the one that follows, we'll turn toward a series of examples that suggest that the BTM is unsatisfactory as a foundational theory of meaning because it massively over-generates expression meaning. In this section, we'll discuss *polysemy over-generation*, which is so named because it occurs when too many meanings are attributed to a single linguistic expression. As it stands, (BTM 1) is guilty of two varieties of polysemy over-generation.

We'll refer to (BTM 1)'s first variety of polysemy over-generation as *vertical*, for reasons that will soon become clear. The BTM relies on the possible effect to which the copying of an item is sensitive, and nearly any such effect will be part of a longer causal chain if it occurs. When the

[14] The BTM's emphasis on the behavioral aspect of use also differentiates it from other use theories that emphasize a mental aspect of use, such as Dummett's (1996) view that it is not just use itself but the *knowledge* guiding speakers' use that gives rise to expression meaning (pp. 104–105), and Horwich's (2004) view that what gives rise to expression meaning is the underived use of sentences containing that expression in drawing *inferences* (pp. 352, 360). While we're comparing the BTM to other work on meaning, it's worthwhile to consider its relationship to work on meaning that in some way shares, or seems to share, the BTM's emphasis on behavior. W.V.O. Quine (1960/2013) treats a key kind of meaning (namely, stimulus meaning) as a matter of dispositions to engage in the behavior of assenting and dissenting to a given sentence in various contexts (pp. 29–30). This differs from the BTM in incorporating actual behavior only as "evidence" for the existence of a disposition, where it is the disposition and not the behavior that actually gives rise to meaning (Quine 1960/2013, p. 37). Bertrand Russell (1980) emphasizes behavior when discussing meaning, but his focus is on a learned tendency to produce a sound when a certain object is present (on the speaker's side) and a tendency to behave as if the object is present when one encounters that sound (on the hearer's side), rather than on effect-sensitive behavior (pp. 14, 61, 66–68). Millikan (2005) offers a foundational theory of meaning that uses her notion of copying (and conventions), but her theory is not entirely behavioral. What is copied to give rise to meaning is a pattern that includes not just the speaker's utterance, but also (in the case of sentences in the indicative mood) the hearer's mental activity of forming a belief (Millikan 2005, pp. 58–59). Brian Skyrms's (2010) work on meaning and the BTM are mutually complementary. The projects are different—Skyrms's (2010) primary aim is not to give a foundational account of the facts in virtue of which items become meaningful expressions, but rather, he investigates the kinds of behavioral mechanisms that could allow expression meaning to spontaneously emerge (p. 1). Building on Skyrms's work, Kevin Zollman (2005) and Elliott Wagner (2009, 2015) have focused on the mechanism of imitation in particular (*cf.* Skyrms 2010, pp. 101–103). This work suggests that the kind of copying behavior which the BTM treats as partially constitutive of expression meaning is also plausible as a mechanism by means of which expression meaning actually emerged. Michael Johnson and Jennifer Nado (2014) offer what appears to be a behavioral foundational theory of meaning, but their view shares Quine's focus on *dispositions* toward linguistic behavior rather than behavior itself (though the particular dispositions to which they appeal are different) (pp. 81–82). Furthermore, Johnson and Nado aim to explain meaning only within the idiolects of particular speakers, so their target is not actually the public notion of expression meaning that the BTM aims to capture.

widespread copying of the item is sensitive to more than one member of that possible chain of effects, (BTM 1) implies that the item has too many meanings. This can happen both with effects that would occur earlier in the causal chain than the one we want to say is meaning-determining, and with effects that would occur further down the chain.

Consider North Americans' behavior with the sound >Disneyland<. They are likelier to produce that sound on occasions on which doing so is likely to get others to involve the Anaheim, California theme park in their activity—their behavior is sensitive to that effect. However, that possible effect is part of a chain of other possible effects. The details are unimportant for our purposes, but we can imagine that the causal chain is something similar to the following: producing >Disneyland< would disturb air particles, then it would vibrate someone's eardrums, then the owner of the eardrums would notice it, then she would understand it, and then she would involve the Anaheim theme park in some kind of activity.

According to our second definition of effect-sensitive behavior, North Americans' behavior with >Disneyland< is sensitive to many or perhaps even all of the effects in the chain. For instance, people are much likelier to produce >Disneyland< in contexts in which it is likely to get someone else to notice the sound, and much less likely to produce it in contexts in which that effect is unlikely (such as when no one else is present, or in a location with too much background noise). Because noticing >Disneyland< is one way of involving the relation of noticing in activity, and there are suitable alternative ways to get people to involve the relation of noticing in activity, (BTM 1) implies that >Disneyland< means the relation of noticing, as well as the theme park in Anaheim.

(BTM 1) causes vertical over-generation in the other direction as well. There is a very likely additional link in >Disneyland<'s causal chain that would come after the effect of getting the hearer to involve the Anaheim theme park in her activity: the hearer (particularly if the hearer is a child) would be likely to feel excitement. In general, people are likelier to utter >Disneyland< in contexts

in which it is likely to cause excitement (*e.g.*, in the presence of young children), and they are less likely to utter it in contexts in which it is less likely to cause excitement (*e.g.*, in the presence of cynical teenagers). In other words, many people's behavior with >Disneyland< is sensitive to the effect of getting others to involve excitement in their activity. People copy each other in using >Disneyland< in this way, and there are plenty of alternative ways to get others to feel excited. However, >Disneyland< obviously does not *mean* excitement; it just means the theme park in Anaheim.[15]

Along with this bidirectional vertical over-generation, there is a second variety of polysemy over-generation that troubles (BTM 1): *horizontal* over-generation. To discuss this form of over-generation, we'll need to consider an entire sentence—specifically, the platitude 'Rome wasn't built in a day.' We can think of the noise we make in uttering the entire sentence as a single sound. Then we'll quickly notice that producing the sound >Rome wasn't built in a day< is a copied, arbitrary way of getting others to involve the city Rome (among other things) in activity. So, (BTM 1) implies that the entire sound >Rome wasn't built in a day< is polysemous and means Rome, the relation of building, the property of being a 24-hour segment of time, *etc.* More generally, (BTM 1) implies that items with complex meanings also mean each part of the complex meaning. This result is unacceptable.[16]

To search for a way for the BTM to avoid all of these varieties of polysemy over-generation, let's return to the link in >Disneyland<'s likely causal chain at which the hearer notices the sound. A step in the right direction is to observe that although people are likelier to produce >Disneyland< in contexts in which doing so is likely to get others to notice the sound, they do this almost exclusively

[15] I am grateful to the participants in the Fall 2013 New York Philosophy of Language Workshop for pressing similar objections.

[16] Ben Lennertz helped me recognize this problem. I used a platitude as my example here, because platitudes are the only sentences that typically get copied directly. In the case of other sentences, the syntactic form and individual words are copied, but generally not the sentence as a whole (*cf.* Millikan 2005, p. 3). This means that non-platitudinous sentences do not satisfy condition (1) of (BTM 1).

when producing >Disneyland< is also likely to get others to involve the Anaheim theme park in their activity. Producing >Disneyland< would still be fairly likely to attract notice even among people who have never encountered that sound before, because speech-like noises draw attention, but we would very rarely bother to do it in those circumstances. In other words, our use of >Disneyland< as a way of getting others to involve the relation of noticing in their activity is *restricted* to cases in which it is also likely to get others to involve the Anaheim theme park in their activity.

We can make a similar observation about the >Rome wasn't built in a day< case. People *are* likelier to produce >Rome wasn't built in a day< in contexts in which it is likely to get others to involve Rome in their activity, but those cases are nearly always ones in which producing that sound is also likely to get others to involve the entire state of affairs in which the construction of that city is not completed within twenty-four hours in their activity. For instance, people in Italy are fairly likely to involve the city Rome in their activity when they hear >Rome wasn't built in a day< (given the similarity of >Rome< to the Italian >Roma<), but they are less likely to involve the entire state of affairs in activity (unless they also speak English). Because we don't tend to produce >Rome wasn't built in a day< unless the entire state of affairs is likely to get involved in the audience's activity, the effect of getting people to involve the city Rome in their activity is restricted by the effect of getting them to involve the entire state of affairs in their activity.

It will be helpful to have an official definition of restriction:

An individual's behavior *B*'s possible effect *G* restricts *B*'s possible effect *F* if and only if the following two conditions are satisfied:

(1) On the occasion on which *B* occurs, *F* is likely to be either:
    (a) part of a causal chain that leads to *G*, or
    (b) part of *G*, where *G* will not occur simply in virtue of *F* occurring.
(2) In the past, when the individual has engaged in behavior of the same kind as *B* in contexts in which *B* was relatively likely to cause *F*, those were nearly always contexts in which the behavior was also relatively likely to cause *G*.

Using this definition, we'll amend the BTM to require that the effects to which speakers' widespread behavior is sensitive be *unrestricted*. The idea is that if we behave in ways that are sensitive to some

effect only (or nearly only) when it is also likely to be a means to or a component of some other effect, the latter effect is much more informative about the nature of our behavior.

This change to the BTM will straightforwardly rule out vertical over-generation cases in which the problematic effect is located earlier in the likely causal chain than the effect that is actually meaning-determining, such as the noticing >Disneyland< case. Clause (1a) in the definition of restriction is crucial here. As we noted above, when people produce >Disneyland< in contexts in which it is likely to get others to notice the sound, that effect is nearly always likely to be part of a causal chain that will lead to getting them to involve the Anaheim theme park in their activity. This entails that getting others to involve the relation of noticing in their activity is a restricted effect of producing >Disneyland<, which means it does not give rise to expression meaning.

Clause (1b) of the definition of restriction will eliminate horizontal over-generation. Involving Rome in activity is part of the larger possible effect of involving the whole state of affairs in activity. Whenever we produce >Rome wasn't built in a day< in contexts in which it is likely to get others to involve Rome in their activity, the context is nearly always such that producing >Rome wasn't built in a day< is also likely to bring about the larger effect of getting them to involve the whole state of affairs in their activity. So, the effect of involving the whole state of affairs in activity restricts the effect of involving Rome in activity, as well as the effects of involving each of the other constituents of the state of affairs in activity. This means that none of the problematic effects are meaning-determining any longer in that case.[17]

[17] An anonymous referee made me aware of a problem about gerrymandered disjunctive effects that is worth mentioning here. When I produce >Disneyland<, I am not only likely to cause my audience to involve the Anaheim theme park in activity, but I am also likely to cause them to involve the Anaheim theme park or the Rosetta stone in their activity. The larger disjunctive effect might seem to restrict the smaller one. Clause (1b) in the definition of restriction precludes this result: if the effect of the audience involving the theme park or the Rosetta Stone in their activity occurs, it will occur only in virtue of the occurrence of the audience involving the theme park in their activity. This contrasts notably with the example about Rome, in which the audience involving the state of affairs in which Rome was not built in a day in their activity would *not* occur just in virtue of the audience involving Rome in their activity.

However, one direction of vertical over-generation still isn't eliminated for >Disneyland<. The effect of getting others to involve the Anaheim theme park in their activity does not restrict the effect of getting them to involve excitement in their activity because involving excitement in activity is not likely to be a part of or a means to the effect of involving the theme park in activity. It's also important to note that the effect of involving excitement in activity doesn't restrict the effect of involving the theme park in activity either, despite the fact that involving the theme park in activity is likely to be part of the means to the excitement. Although we often produce >Disneyland< when doing so is likely to cause excitement, we also produce that sound in plenty of contexts in which involving the Anaheim theme park in activity is a likely effect but excitement is not. Not everyone gets excited about Disneyland, but we still produce >Disneyland< as a way of getting such people to involve the theme park in their activity, perhaps to criticize Disneyland.

The excitement case shows that the BTM cannot allow all of the unrestricted effects to which our behavior with some item is sensitive to be meaning-determining. More specifically, it shows that for the BTM, the meaning-determining effect should be the *first* unrestricted possible effect to which our behavior with an item is sensitive. We start at the beginning of the chain of effects to which the behavior is sensitive and proceed until we find an unrestricted effect that can be produced in suitable alternative ways, and that effect is what determines the item's meaning. This rules out our remaining polysemy over-generation case. >Disneyland<'s likely effect of getting others to involve excitement in their activity, though unrestricted, is not the *first* unrestricted effect to which the behavior is sensitive. It itself is likely to be the result of another unrestricted effect: getting others to involve the Anaheim theme park in activity. Involving the Anaheim theme park in activity, on the other hand, is the first unrestricted effect to which the behavior is sensitive, so it determines the sound's meaning.

We are now ready for an improved version of the BTM, which differs from (BTM 1) only by the addition of condition (3):

(BTM 2) A type of item *I* means object or relation *O* within a group in virtue of all of the following conditions being satisfied:

(1) Within the group there is widespread, interconnected copying of producing *I* as a way of getting others to involve *O* in their activity.
(2) There is at least one other way of getting others to involve *O* in their activity that is at least approximately as conducive to that effect and at least approximately as accessible to the group as producing *I*, independent of the dominance producing *I* has gained due to being copied.
(3) When group members produce *I* as a way of getting others to involve *O* in activity, it is typically the case that getting others to involve *O* in activity is an unrestricted possible effect of producing *I* which is not itself likely to be the result of another unrestricted possible effect.

Condition (3) accommodates genuine polysemy by allowing for a "tie" for the status of first unrestricted possible effect. When an expression is not polysemous, there will be only one possible effect that satisfies condition (3). 'Typically' in condition (3) is meant in a fairly light sense. The BTM doesn't need to require that meaning-determining effects are absolutely *never* restricted; rather, it just needs to be the case that overall throughout the group, they are unrestricted most of the time. Importantly, condition (3) also allows (BTM 2) to be entirely behavioral: it appeals just to speakers' behavior in circumstances in which various effects on the hearer (including, perhaps, some mental effects) are possible or likely, with no appeal to the actual occurrence of mental phenomena.

5. Increased Conduciveness

(BTM 1) required modification because it implied that linguistic expressions had too many meanings, but now we'll see that (BTM 2) still implies that things that clearly *aren't* linguistic expressions are meaningful. We'll refer to this problem as *misplacement over-generation* because it misplaces expression meaning. Misplacement over-generation is equally problematic, so the BTM needs another modification.

Imagine a group of office workers who all go on diets to lose weight from time to time. One of them, David, reaches the weight recommended by his physician and ends his diet. David isn't particularly kind, and after ending his own diet he leaves a cupcake out in the break room in order to get his co-workers to violate their diets. It works: one of his dieting co-workers eats the cupcake.

Others in the office eventually find out about David's trick, and many of them occasionally copy him in leaving cupcakes as a way of getting their co-workers to violate their diets. The behavior has alternatives: there are other ways the co-workers could achieve the same result (*e.g.*, by leaving pie or cookies in the break room). The prior and partial possible effects to which the behavior is sensitive are restricted: for instance, the co-workers rarely try to get people to involve cupcakes in their activity when it is not likely to lead them to violate their diets. And there isn't any further possible effect that restricts the effect of getting people to violate their diets. So, (BTM 2) implies that within this group, cupcakes are linguistic expressions that have the relation of violating a diet as their meaning. This is a clear misplacement of expression meaning.

To avoid this result, it might be tempting to manipulate the definition of 'item' to exclude cupcakes. However, such a definition of 'item' would likely also exclude objects—such as piles of stones left to indicate the direction of hiking trails—that actually seem like good candidates for expression meaning. And even if we did exclude cupcakes from counting as items, the gesture of leaving a cupcake could lead to the same over-generation problem.

But there's a more natural way to rule out this case: the BTM can require that when an item becomes a meaningful expression, producing that item becomes significantly more conducive to whatever effect it is copied as a way of achieving. So, in our example, cupcakes would not be meaningful expressions because leaving them in the break room did not become likelier to get people to involve the relation of violating a diet in their activity as the practice with the cupcakes became widespread. In fact, if anything, the practice would probably become less effective over time as people got tired of eating cupcakes. Producing >apple<, on the other hand, went from being not conducive to getting people to involve the property of being a fruit produced by a tree in the rose family in their activity, to being extremely conducive to that effect. On this picture, when an item becomes a meaningful expression in a human natural language, it gains a new status in quite a strong sense:

through copying, producing that item becomes very conducive to an effect that it was unlikely to bring about before.

We are now ready for our final version of the BTM, which differs from (BTM 2) only by the addition of condition (4):

(BTM 3) A type of item *I* means object or relation *O* within a group in virtue of all of the following conditions being satisfied:

(1) Within the group there is widespread, interconnected copying of producing *I* as a way of getting others to involve *O* in their activity.
(2) There is at least one other way of getting others to involve *O* in their activity that is at least approximately as conducive to that effect and at least approximately as accessible to the group as producing *I*, independent of the dominance producing *I* has gained due to being copied.
(3) When group members produce *I* as a way of getting others to involve *O* in activity, it is typically the case that getting others to involve *O* in activity is an unrestricted possible effect of producing *I* which is not itself likely to be the result of another unrestricted possible effect.
(4) Producing *I* became significantly more conducive to the effect of getting others to involve *O* in their activity due to the copying that caused it to become widespread.

In condition (4), the "due to the copying" portion should be understood to imply that the copying must be a *proximate* cause of the increase in conduciveness. Moreover, like condition (3), condition (4) is compatible with the BTM's aim to be entirely behavioral: it appeals just to an increase in the likelihood of some (possibly mental) effect on hearers, and not to the actual occurrence of any mental phenomena.

So, according to this version of the BTM (which will be the final version for our purposes), an item becomes a meaningful linguistic expression when production of it becomes some group's widespread, copied way of getting others involved with some part of the world, where getting others involved with that part of the world is the first unrestricted possible effect at which group members' behavior with the item actually aims, and where producing that item became a much better way to get others involved with that part of the world as it gained its new status.

It's important to note that without the word 'significantly' in condition (4), (BTM 3) would still over-generate meaning. Imagine a group of people who sometimes mock others' sneezes by mimicking sneezing. They copy each other in mimicking sneezes in one particular way from among

the many slightly different ways that would be equally recognizable, and getting others to involve sneezing in activity is the first unrestricted possible effect to which these performances are sensitive. Once their practice becomes widespread, there might well be a small increase in how conducive their way of mimicking sneezes is to getting people to involve the property of sneezing in their activity, but that doesn't seem to make it the case that the group's way of mimicking sneezing is a meaningful linguistic expression, a part of their language.

There doesn't seem to be any hope of specifying a firm cutoff for the absolute quantity of increase in conduciveness toward the relevant effect that is required. By requiring that the increase be significant, (BTM 4) is requiring that the increase in conduciveness not be trivially small in comparison to that type of item's preexisting conduciveness to that possible effect. Producing >apple< was initially not at all conducive to getting people to involve in activity the relation it now means, so gaining a fairly small amount of added conduciveness was significant in that case. On the other hand, the performance used to mimic sneezing was already highly conducive to the effect of getting people to involve sneezing in activity, due to the high degree of natural resemblance. The small amount of increased conduciveness it might gain from becoming the group's widespread way of mimicking sneezing is trivial in comparison to the large amount of conduciveness it already had.

6. Semantic Deference

Before concluding, I'd like to discuss one final concern about the BTM. One might worry that the BTM will run into trouble with accounting for *semantic deference*. Loosely, semantic deference occurs when members of a larger linguistic group defer to some subset of experts with respect to the meaning of certain words (common examples include medical or scientific terms, such as 'cancer' or 'acid'). One might think that this sort of deference is at least a factor in determining the meaning of those words in the larger group, and that this influence on expression meaning might be difficult to capture

without bringing in participants' mental states. If that's correct, then even though the BTM may succeed in its aim to be entirely behavioral, it will give the wrong verdict about the meaning of certain terms—those whose meanings are determined by more than just behavioral factors, due to semantic deference.[18]

In my view, there are actually three importantly different phenomena that can plausibly be characterized as semantic deference. I'll describe each of them in turn, arguing that the BTM correctly implies that only one of the three leads to changes in expression meaning within the larger group.

One phenomenon that might be called 'semantic deference' occurs when members of a larger linguistic group change their widespread usage of some term to match the usage of a subgroup of experts. For example, if members of the U.S. medical community determine that a subtle shift in which property we use >cancer< to mean is desirable (perhaps because of a new discovery about the shared etiology of certain pathological properties) and then they make the shift in their own usage, members of the larger linguistic group might be influenced by the medical community to alter their usage, too. There it seems undeniable that the members of the larger linguistic group have deferred to the medical community, and that this semantic deference has resulted in a change to the public meaning of >cancer< within the larger linguistic group.

As is probably already clear, the BTM can account quite easily for this form of semantic deference. Once people have copied the experts' changed behavior to an extent that makes the new usage is widespread within the larger linguistic group, the BTM's verdict will be that the meaning of >cancer< within the broader linguistic group has changed. Thus, there is one form of semantic deference that clearly does produce a change in expression meaning and which the BTM seamlessly accommodates.

[18] I am grateful to an anonymous referee for bringing this concern to my attention.

Interestingly, there is also a BTM-inspired way to describe the expert subgroup's particularly influential status over time, by noting that the behavior of members of the subgroup is copied more frequently (and leads to longer chains of further copying) than is the behavior of other members of the larger linguistic group. Moreover, we could introduce a BTM-inspired notion of the larger group actively tracking the subgroup's usage: we can observe that as the experts shift their usage of >cancer< over time, members of the larger group tend to shift their usage as well. This notion of tracking could rely on the simple notion of differential likelihood of behavior already used in our notion of effect-sensitive behavior.

A second kind of phenomenon that seems intuitively categorizable as semantic deference involves changes in usage that are situation-specific rather than widespread within the larger linguistic group. Here, consider the common colloquial usage of >acid<. In some English-speaking groups, there is widespread use of >acid< to get people to involve in their activity the relation of being able to dissolve proteins and other commonly encountered substances. For instance, people speak of "battery acid" even in the case of alkaline batteries, and they typically don't describe something as gentle as vinegar as an acid. And yet, there is a subgroup of experts (namely, chemists) who use >acid< to get people to involve in their activity the property of being an aqueous solution with a pH below seven. Members of the larger linguistic group display a kind of semantic deference toward chemists—if we take a chemistry course, for instance, we adjust our usage of >acid< to match chemists' usage while we're in class. And if I called a substance leaking out of an alkaline battery 'acid' in the presence of a chemist who corrected me, I would accept the correction. These corrections from experts do not result in changes to widespread usage—even those who have taken a chemistry class often go on to use >acid< when talking about the substance inside alkaline batteries and not when talking about vinegar. Yet, this phenomenon certainly seems categorizable as semantic

deference: members of the larger group are clearly deferring to members of the expert subgroup about the meaning of >acid<.

The BTM's verdict about this kind of case will be that the deference exhibited by members of the larger linguistic group does not affect the meaning of >acid< in that larger group, because it does not change widespread behavior. As always, the BTM holds that widespread behavior is what determines expression meaning. The question, then, is whether this is a problem for the BTM—whether a good foundational theory of meaning *ought* to categorize this kind of deference as resulting in a change in expression meaning within the larger group.

Leaving the BTM aside for the moment, I think it's quite plausible that this second form of semantic deference does not result in a change in expression meaning within the larger group. Expert subgroups have a kind of heightened social status within the larger group. Thus, in a situation in which a member of the expert subgroup's expertise is relevant, social norms dictate all sorts of deference to that individual, beyond just semantic deference: in a chemistry class, we follow instructions from the teacher about which substances to combine and how to do so, along with accepting her corrections about how to use technical terms. Or in non-classroom situations in which chemicals are in play and a chemist is present (such as after a spill of industrial cleaning supplies), we might take on the same sort of broadly deferential stance. But this deference is limited, ending when the situation (such as chemistry class) ends. And in fact, in the case of >acid<, it's easy to see why the deference doesn't extend beyond these very specific situations to affect widespread usage in the larger group: the current widespread use of >acid< is more useful to us than the chemists' usage, in ordinary contexts. I don't very often need to know the pH of some substance that is in front of me, but I *do* frequently need to know whether it could dissolve my skin or perhaps damage the finish on my table if I spill it. Thus, it strikes me as quite plausible to say, as the BTM does, that this sort of

situation-specific semantic deference does not affect the meaning of a linguistic expression within the larger group.

A third kind of phenomenon that might be categorized as semantic deference involves only *attitudes* of deference. Here I'll turn to a definition of semantic deference offered by Diego Marconi (2012):

> Minimally, semantic deference can be characterized as follows:
> 1. Ordinary, non-expert speakers know that word W has an expert usage that may differ from their own;
> 2. They believe that such expert use is the correct one;
> 3. They assume that their own use of W is consistent with the expert's;
> 4. However, they are prepared to amend it if it is shown to be inconsistent with the expert's use (pp. 273–274).

This definition of 'semantic deference' would presumably apply to many cases in which usage actually changes, but we've already discussed how the BTM can accommodate the change in expression meaning in such cases. So, I'd like to focus on cases in which Marconi's four conditions are satisfied, but the widespread usage in the broader linguistic group does not match the experts' usage. For instance, imagine that laypeople in the U.S. are aware that there is an expert usage of >cancer< (satisfying condition (1)), they believe that the experts' usage is correct (satisfying condition (2)), they believe that their usage matches the experts' usage (satisfying condition (3)), and they are disposed to change their usage if confronted with an inconsistency with the experts' usage (satisfying condition (4)). But imagine that most laypeople's linguistic behavior is *not* consistent with the experts' usage, and they never encounter any evidence of this disparity, so their behavior never changes. Marconi's definition would categorize this situation as involving semantic deference.

Again, the BTM will render the verdict that the meaning of a linguistic expression within the larger linguistic group is determined by widespread behavior within that larger group. So, the BTM implies that in the scenario just described, >cancer< has a different meaning among people in the U.S. as a whole than it does within the medical community, despite the laypeople's widespread deferential

attitudes toward the experts. The question, then, is whether these mere deferential attitudes *should* be seen as, on their own, changing the meaning of >cancer< within the larger linguistic group.

My view is that this kind of situation is plausibly described as one in which members of the larger linguistic group are *primed* to defer, rather than one in which they actually *are* deferring, and thus there is nothing implausible about claiming that expression meaning is not affected by the deferential attitudes alone. Due to a false belief (namely, their belief that their behavior matches the experts' usage), laypeople in our example wrongly *think* they are deferring to the experts in their behavior with >cancer<, and due to their disposition to change their behavior to match the experts if they became aware of the disparity, they are primed to actually defer. For someone who is at all sympathetic to the BTM, having to characterize such a case as one in which people are merely primed to defer and wrongly think they are deferring will probably not seem like too bitter a pill to swallow. Furthermore, nothing about the BTM prevents us from describing what *is* deferential about the situation: we can say that there are deferential attitudes, even though no one is actually deferring.

Another worry in the neighborhood, however, might be that even if the mental states in Marconi's definition of semantic deference are not sufficient to change expression meaning on their own, perhaps the BTM still forces us into an impoverished account of semantic deference by leaving those mental states out of the picture. Even if we accept that the mere presence of deferential attitudes can only create a situation in which people are *primed* to defer, still those attitudes do often accompany actual deferential behavior when it occurs, and they do seem to play an important role in such situations.

It is, in fact, perfectly compatible with the BTM to say that mental states along the lines of those in Marconi's definition *are* fairly likely to be present when people engage in deferential linguistic behavior, and that they are likely to guide and motivate that behavior. But as was the case for the intentions that guide ordinary linguistic behavior, the BTM focuses only on the behavioral aspect as

what gives rise to expression meaning. All of this is compatible with saying that the behavioral phenomena that give rise to expression meaning are embedded in a structure of other social practices, some of which are partially grounded in mental phenomena. For instance, our linguistic conventions (themselves entirely behavioral) can interact with social norms about the relative status of various groups (as in our >acid< example) or perhaps with our group's conventional beliefs about how our usage relates to expert usage (as in the most recent >cancer< example) to make up the overall structure of our linguistic practice, which is much broader and richer than the part that determines expression meaning. So, I do think that one could endorse the BTM while also doing justice to the role of mental states, such as the ones involved in the kind of deferential attitudes Marconi describes, in linguistic practice more broadly. The present paper will not accomplish this task, because its focus is just on exploring a possible account of what determines expression meaning, but that is no reason to think it could not be done.

I'd also like to highlight a virtue of the BTM-friendly approach to semantic deference that we've outlined: it categorizes situations in which people are influenced by some subgroup without being aware of being so influenced as semantic deference. Members of the larger linguistic group might not have any opinion about whether the expert usage of terms such as 'cancer' and 'acid' should have any kind of dominance, and they may hold no beliefs or assumptions about how similar their behavior is to expert usage. But nonetheless, they might be unconsciously influenced by signals of dominance (such as the white coats and peremptory manner of physicians) to defer to those individuals, without any awareness that they are deferring. If this influence results in widespread changes in linguistic behavior, it seems to me that it should still be counted as semantic deference that affects expression meaning, even in the absence of the mental states on Marconi's list.

So, the BTM seems to have a plausible response to the concern about semantic deference. The theory seamlessly accommodates changes in expression meaning for the kind of semantic

deference that most clearly produces such changes, while still leaving room for conversations about a role for deferential attitudes and the norms that give rise to them in our understanding of linguistic practice more broadly.

7. Conclusion

The primary aim of this paper has been to explore the plausibility of a foundational theory of meaning that departs from the dominant trend in the literature by eschewing mental states of all kinds, focusing instead on language users' observable behavior as what gives rise to expression meaning. Effect-sensitive behavior played a key role as we developed the BTM, but it quickly became clear that the notion of effect-sensitive behavior is quite permissive—that is, many instances of behavior are sensitive to a wide array of effects. This permissiveness led to serious concerns about the BTM's adequacy as a foundational theory of meaning, due to massive over-generation of expression meaning in (BTM 1). However, we saw that it's possible to modify the BTM to avoid that over-generation, and that the BTM has resources to respond to a concern about its ability to accommodate semantic deference.

Before wrapping up, I want to note as a matter of clarification that the BTM does not present any of its conditions as individually necessary for expression meaning.[19] Rather, (BTM 3) aims to provide a particularly interesting set of jointly sufficient conditions for expression meaning: the ones in virtue of which the linguistic expressions in actual human natural languages are meaningful. (BTM 3) aims to answer the particular question with which this paper began: the question of what makes it

[19] In fact, there are cases that suggest that conditions (1) and (2), as stated in this paper, are not necessary: expression meaning could arise among beings that have evolved to have an entirely innate communication system (Peacocke 1976, pp. 168–169), or among beings that are endowed by a chance event (such as radiation exposure) with a single communication system from which they are then unable to deviate (Armstrong 2016a, p. 103). I am grateful to Daniel Harris for discussion of these issues.

the case that certain arbitrary sounds and other items that humans produce have the feature of public, group-level meaningfulness (*cf.* Bennett 1976, pp. 22–23).

It may seem that at least conditions (3) and (4) of (BTM 3) must be necessary conditions for expression meaning, because they were introduced to *rule out* cases that (BTM 1) and (BTM 2) wrongly categorized as involving expression meaning. However, all this means is that conditions (3) and (4) are necessary features of the particular set of sufficient conditions in (BTM 3). In other words, all that an advocate of (BTM 3) is committed to is that conditions (3) and (4) are necessary to take us from the insight that copying behavior is the likely explanation for the ability of human linguistic expressions to overcome the obstacle of arbitrariness, to a set of conditions that really are jointly sufficient for expression meaning.

Much work remains to be done in relation to exploring the plausibility of the BTM. As we acknowledged early on, the BTM needs to be expanded beyond morphemes that have objects or relations as their meanings, which will require engagement with empirical work on the semantics of particular kinds of expressions. Further expansion is also needed to take the BTM beyond individual morphemes, to the level of sentences. Additionally, because the BTM relies on effects to which linguistic behavior throughout an entire group is observably sensitive, it is easy to be concerned that it will lead to widespread indeterminacy of meaning. The answer to this worry will be complex and is likely to consist partly of an acknowledgement that there truly *is* some indeterminacy, but that this is a discovery about the nature of expression meaning rather than a problem for the BTM. Much more remains to be said about these issues, and I look forward to undertaking those projects in future work.

Acknowledgements:

I am grateful to the following people for helpful conversations related to this paper: Peter Graham, Daniel Harris, Ruth Millikan, Michael Nelson, John Perry, John Ramsey, Indrek Reiland, Howard Wettstein, Larry Wright, and members of a work-in-progress group at the University of California, Riverside. I'm also grateful for feedback from several anonymous referees and from participants in the Fall 2013 New York Philosophy of Language Workshop, 2014 Midsouth Philosophy Conference, 2014 SoCal Philosophy Conference, 2015 Central APA Meeting, and the 2018 *Topoi* Conference on Foundational Issues in Philosophical Semantics. This paper contains and builds on material from my dissertation, *Conventions and Expression Meaning* (University of California, Riverside, 2016).

References

Armstrong J (2016a) Coordination, triangulation, and language use. *Inquiry* 59(1):80–112

Armstrong J (2016b) The problem of lexical innovation. *Linguistics and Philosophy* 39(2):87–118

Bennett J (1976) *Linguistic behaviour*. Cambridge University Press, Cambridge

Burge T (1975) On knowledge and convention. *Philosophical Review* 84(2):249–255

Burgess A & Sherman B (2014) Introduction: A plea for the metaphysics of meaning. In: Burgess A & Sherman B (eds) *Metasemantics: New essays on the foundations of meaning*. Oxford University Press, Oxford

Davidson D (1973) Radical interpretation. *Dialectica* 27(3):313–328

Davis W (2003) *Meaning, expression, and thought*. Cambridge University Press, Cambridge

Davis W (2005) *Nondescriptive meaning and reference: An ideational semantics*. Oxford University Press, New York

Dummett M (1996) What do I know when I know a language? In: *The seas of language*. Oxford University Press, Oxford

García-Carpintero M (2012a) Foundational semantics i: Descriptive accounts. *Philosophy Compass* 7(6):397–409

García-Carpintero M (2012b) Foundational semantics ii: Normative accounts. *Philosophy Compass* 7(6):410–421

Gilbert M (1989) *On social facts*. Routledge, London

Grice P (1989) *Studies in the way of words*. Harvard University Press, Cambridge, MA

Hawthorne J (2007) Craziness and metasemantics. *The Philosophical Review* 116(3):427–440

Horwich P (1998) *Meaning*. Oxford University Press, New York

Horwich P (2004) A use theory of meaning. *Philosophy and Phenomenological Research* 68(2):351–372

Israel D, Perry J, & Tutiya S (1993) Executions, motivations, and accomplishments. *The Philosophical Review* 102(4):515–540

Johnson M & Nado J (2014) Moderate intuitionism: A metasemantic account. In: Booth AR & Rowbottom DP (eds) *Intuitions*. Oxford University Press, Oxford

Kaplan D (1989) Afterthoughts. In: Almog J, Perry J, & Wettstein H (eds) *Themes from Kaplan*. Oxford University Press, New York

Lewis D (1969/2002) *Convention: A philosophical study*. Blackwell, Oxford

Lewis D (1975) Languages and language. In: Gunderson K (ed) *Language, mind and knowledge: Minnesota Studies in the Philosophy of Science*, vol 7. University of Minnesota Press, Minneapolis

Loar B (1976) Two theories of meaning. In: *Truth and meaning: Essays in semantics*. Oxford University Press, London

Marconi D (2012) Semantic normativity, deference and reference. *Dialectica* 66(2):273–287

Marmor A (2009) *Social conventions: From language to law*. Princeton University Press, Princeton

Miller S (2001) *Social action: A teleological account*. Cambridge University Press, Cambridge

Millikan R G (2005) *Language: A biological model*. Oxford University Press, Oxford

Peacocke C (1976) Truth definitions and actual languages. In: Evans G & McDowell J (eds) *Truth and meaning: Essays in semantics*. Oxford University Press, London

Quine WVO (1960/2013) *Word and object*. MIT Press, Cambridge, MA

Russell B (1980) *An inquiry into meaning and truth*. George Allen & Unwin, London

Schiffer S (1972) *Meaning*. Oxford University Press, London

Simchen O (2017) Metasemantics and singular reference. *Noûs* 51(1):175–195

Skyrms B (2010) *Signals: Evolution, learning, and information*. Oxford University Press, Oxford

Speaks J (2015) Theories of meaning. In: *The Stanford encyclopedia of philosophy, 2015 Edition*. http://plato.stanford.edu/archives/fall2014/entries/meaning/. Cited 24 July 2019

Stalnaker R (2003) Reference and necessity. In: *Ways a world might be: Metaphysical and anti-metaphysical essays*. Oxford University Press, New York

Stotts MH (2017) Walking the tightrope: Unrecognized conventions and arbitrariness. *Inquiry* 60(8):867–887

Stotts MH (forthcoming) Toward a sharp semantics/pragmatics distinction. *Synthese*

Wagner E (2009) Communication and structured correlation. *Erkenntnis* 71(3):377–393

Wagner E (2015) Conventional semantic meaning in signalling games with conflicting interests. *The British Journal for the Philosophy of Science* 66(4):751–773

Williams JRG (2007) Eligibility and inscrutability. *The Philosophical Review* 116(3):361–399

Wittgenstein L (1953/2001) *Philosophical investigations: The German text, with a revised English translation.* Anscombe GEM (trans) Blackwell Publishing, Malden

Zollman KJS (2005) Talking to neighbors: The evolution of regional meaning. *Philosophy of Science* 72:69–85