# Vague Representation

Agustín Rayo
arayo@ucsd.edu

June 1, 2005

Contemporary discussions of vagueness tend to suffer from two important limitations. The first is a lack of generality. More often than not, they focus exclusively on vagueness in public language, even though vagueness afflicts representation in general. And more often than not, it is unclear how the proposal is supposed to extend to other forms of representation. The second limitation arises when it comes to the task of explaining what vague representation consists in. Proposals tend to fall into one of two categories: those that are forced to countenance unexplained boundaries, and those that are limited in their ability to convey useful information about the nature of vague representation. Let me explain.

A proposal might be forced to countenance sharp boundaries even if it doesn't wear them on its sleave. Here are some examples: (*i*) vague representation is context-sensitive, but there are sharp boundaries relative to any given context; (*ii*) vague representation is a matter of degree, but for any given degree there is a sharp boundary between cases to which a representation applies to that degree and cases to which it doesn't; (*iii*) although there is no sharp boundary between cases to which a representation applies and cases to which it doesn't, there is a sharp boundary between cases to which a representation *definitely* applies and cases to which it doesn't; (*iv*) although there is no sharp boundary between cases to which a representation applies and cases to which it doesn't and no sharp boundary between cases to which it definitely applies and cases to which it doesn't and no sharp boundary between cases in which it definitely definitely applies and cases to which it doesn't, and so forth, there is a sharp boundary between cases to which a representation definitely$^*$ applies and cases to which it doesn't, where definite$^*$ application is an application that is definite, definitely definite, definitely definitely definite, and so forth (into the transfinite, if necessary).

The trouble with postulating sharp boundaries is that something important tends to be left unexplained. There tends to be no explanation, not even in principle, about why the sharp boundary one has postulated isn't in some very proximal but distinct location. Leaving important matters irreparably unexplained doesn't make a proposal false, but it does make the search for an alternative more urgent.

It is at this point that the *standard move* is usually invoked. One invokes the standard move when one constructs one's theory of vagueness in the following three steps: one begins with a theory that elucidates the transition between cases to which a vague representation applies and cases to which it doesn't by postulating sharp boundaries of some kind or another; next, one argues that one's metalanguage is vague in certain crucial respects; finally, one claims that an effect of the newly postulated vagueness is that one's proposal is not committed to sharp boundaries after all. The problem with the standard move is that it results in proposals that are limited in their ability to convey useful information about the nature of the transition. Consider an analogy. You wish to know whether your friend Harry has left the party and I say 'all the bald people have left the party'. If Harry is a clear case of baldness, then you can use my utterance to learn that Harry has left the party. But if Harry is a borderline case of baldness, then you are unable to use my utterance to learn whether Harry has left the party. In the latter scenario, my utterance is objectionable as a proposal about Harry's whereabouts. But the complaint is not that the proposal is *false*. Nor is it that the use of vague language in a proposal about Harry's whereabouts is somehow unacceptable (after all, the same sentence was uttered in the former scenario and there you got the information you needed). What is objectionable is that the vagueness gets in the way: it interferes with your ability to use my utterance to determine whether or not Harry is at the party. Similarly, the complaint about the standard move is not that the resulting proposal is *false*. Nor is it that the use of a vague metalanguage is somehow unacceptable (it is presumably impossible to come up with a *non*-vague metalanguage, after all). What is objectionable is that the vagueness gets in the way: it interferes with one's ability to use the proposal to learn about the nature of vague representation.

Consider, for example, the statement that a certain vague representation has cases of definite application, cases of neither definite application nor definite non-application and cases of definite non-application. By arguing that 'definitely' is vague in the right sort of way, one could claim that such a statement is not committed to the postulation of a sharp boundary between, e.g. the cases of definite application and the rest. But the result of making such a move is *not* that the statement can now be used to elucidate the transition between cases to which the representation definitely applies and cases to which it doesn't without requiring the postulation of sharp boundaries. The result is rather that one is unable to use the statement to learn what the transition is like. In particular, one is unable to use the statement to learn whether or not there are any sharp boundaries. And by failing to learn whether or not there are any sharp boundaries one fails to learn something important about the nature of vague representation. In general, the effect of applying the standard move is that the resulting proposal fails to convey useful information about the most interesting parts of the transition, since the blind-spots will be located at just the places where the proposal would be forced to countenance sharp boundaries were it not for the standard move.

It seems to me that if an account of vagueness is to have any hope of success it must find a way of overcoming these two limitations. On the one hand, it must focus on the phenomenon of vague representation in general, and treat the phenomenon of vagueness in public language as a special case. On the other, it must be uncompromising in what it has to say about the nature of vague representation: no unexplained boundaries, no obstructive vagueness in the metalanguage. The aim of this paper is to defend a proposal meeting these constraints.[1]

# 1 Preliminaries

## 1.1 The game show

You and I are contestants in a game show. The show's producers have placed Harry in a room with several other men. We will win a prize if we are able to single out Harry from the rest. I am told who Harry is, and you must do the singling out. We are to proceed by elimination: I am to send you a message, and you are to ask some (but not all) of the men to leave the room on the basis of the message you have been sent. After a few rounds of elimination, we should be left with precisely one man in the room. Our prize money is reduced every round, so it is in our interests to single out Harry in as few rounds as possible.

To make things interesting, the show's producers have placed an important limitation on the sorts of messages I am allowed to send you. All I have is a stack of photographs. At each stage in the elimination I am to hand you a photograph from my stack, and you must decide which of the men are to leave the room by deciding which of the various ways of partitioning the group is rendered most salient by the photograph you have been given. No further communication between us is allowed. Nor are we allowed to come to an agreement about how the photographs are to be interpreted beforehand.

If one of the photographs in my stack turned out to be a photograph of Harry, I could hand you that photograph, and hope that it would lead you to eliminate everyone but him. Unfortunately, the show's producers have taken care to make things difficult for us. None of the photographs in my stack are photographs of men in the room. So I must find less direct ways of rendering a particular partition of men in the room salient to you.

We will sometimes be aided by the circumstances. Suppose, for example, that there are only two men left in the room. One of them is clearly bald and the other is clearly not bald, but they are otherwise very much alike. If one of the pictures in my stack is a picture of a man who is clearly bald, I might hand it to you and hope that it will lead you to eliminate the full-haired man from the room.

We will sometimes be able to use the picture of the bald man to make successful eliminations in less favorable circumstances. Here are some examples:

1. There are many men in the room, but they are sharply divided into two sharply demarcated categories: men in the first category are all clearly bald and men in the second category are all clearly non-bald. I hand you the picture of the bald man. It seems to you that the most salient way of partitioning the group in light of the picture I have given you is by leaving everyone who is clearly bald in the room and asking everyone else to leave.

[1] I am deeply indebted to a number of texts that are not referenced directly in the discussion below. I should especially mention Lewis (1969), Wright (1976), Stalnaker (1979), Sainsbury (1990), McGee and McLaughlin (1994), Williamson (1994), Laurence (1996), Graff (2000) and Dorr (2003).

2. There are many men in the room, but they can be divided into two sharply demarcated categories: men in the first category are all clearly non-bald and men in the second category are all borderline cases of baldness. I hand you the picture of the bald man. It seems to you that the picture renders salient two distinct partitions: on the one hand, a partition according to which everyone should be asked to leave the room; on the other, a partition according to which the borderline cases of baldness should be asked to stay and the clear cases of non-baldness should be asked to leave. It is common knowledge that asking everyone to leave would be pointless, since that would ensure that we not win the prize. So you ask the clear cases of non-baldness to leave and the borderline cases to stay. (A converse situation in which men in the first category are all clearly bald and men in the second category are all borderline cases of baldness is analogous.)

3. For every $n$ between 0 and 2000, the room contains a man with precisely $100n$ hairs on his head, suitably distributed. (We may either assume that you have only your powers of observation to estimate the number of hairs on people's heads, or that each man is holding a card stating how many hairs he has on his head.) I hand you the picture of the bald man. It seems to you that no particular partition of the group is rendered salient by the picture you have been given. It is common knowledge, however, that you must carry out some division or other. You therefore make an arbitrary choice. You do your best to ensure that nobody who is asked to stay is clearly less bald than someone who is asked to leave. You also try to strike a balance between making the elimination worthwhile by asking enough people to leave and making it safe by asking enough people to stay. Although your particular choice of partition is arbitrary, it is sensible in light of the considerations at hand.

When the circumstances are right, the photograph-method can be very successful in allowing us to narrow down the group of men in the room. But there are limits to what it can achieve. Cases in which you are forced to make an arbitrary decision—such as 3 above—are cases in which I have no control over the exact partition that will result from my handing you the photograph. I must be prepared to tolerate some deviation from the partition that I would have chosen myself.

## 1.2 Semi-principled representation

The photograph-method of the preceding section has an important feature: whether or not my handing you the photograph of the bald man somehow encodes a *rule* for classifying men in the room, it is not a rule you gain access to when I hand you the photograph. As evidenced by the case in which the room contains a man with precisely $100n$ hairs on his head for every $n$ between 0 and 2000, the photograph does not generally enable you to classify the men in the room without making a *decision* about how to use the information at your disposal. Your decision will not be entirely unprincipled: it will be a sensible one in light of the information you have at your disposal. But it will be a decision nonetheless.

I shall say in such cases that classification is achieved through a *semi-principled* classification-method. The difference between a rule-based classification-method and a semi-principled classification-method is akin to the difference between a sign forbidding access to the forest with a map detailing the forest's dangers. From the sign you are able to extract a rule for classifying the relevant possibilities—entering the forest and not entering the forest—but from the map you are not. There is a certain kind of *incommensurability* between the information supplied by the map and the possibilities under consideration. In order to use the map to classify the possibilities you must make a decision about how to use the information that has been made available to you. Under the right circumstances, the decision might be relatively straightforward—for instance, the second sign might indicate that the forest contains poisonous spiders and you might have a strong preference to avoid places containing poisonous creatures. But in the general case your decision will involve balancing off considerations of different sorts—e.g. your fear of spiders versus your desire for adventure—and there may well turn out to be no way of settling on a particular classification without making some arbitrary choices along the way.

There is a similar kind of incommensurability between the information supplied by my handing you the photograph of the bald man and the various ways of classifying the men in the room. It is because there is no predetermined way of bridging this gap that you are forced to make a decision. You have considerations of various sorts at your disposal (e.g. your knowledge of the game-show's setup, the extent of your aversion to risk, your beliefs about what we are both likely to regard as salient, your beliefs about my decision-making process). But more often than not, the considerations at your disposal will fail to provide a single piece of advice about how to classify the men in the room. Different ways of weighing distinct considerations and different ways of making guesses about information you lack are likely to yield different results. And even when different ways of balancing off the various considerations point broadly in the same direction, there may be more than one way of classifying the men in the room which seems roughly satisfactory to you. So even though your classification is not unprincipled, there are a number of places in which it may call for arbitrary decisions: there may be arbitrariness in deciding which considerations should be taken into account, arbitrariness in deciding how the various considerations are to be balanced off against each other, arbitrariness in deciding how to apply the considerations themselves, arbitrariness in your guesses about missing information and arbitrariness in your particular choice of classification after the various considerations have been taken into account.

It is important not to confuse a semi-principled classification-method with a classification-method based on an *incomplete* rule. An incomplete rule mandates certain choices and leaves others open. In the case of a semi-principled classification-method, on the other hand, there is no such thing as a mandated choice (which is not to say, of course, that choices must be entirely unprincipled).

### 1.3 Tolerance

Let me introduce some additional notation. Say that a semi-principled classification-method $\mathcal{M}$ displays *tolerance* with respect to parameter $\Pi$ (as used by subject $S$) if there is some range of possibilities $P_\Pi$ such that each of the following four conditions obtains:

1. The possibilities in $P_\Pi$ differ with respect to parameter $\Pi$, but are otherwise similar in all relevant respects.
2. The possibilities in $P_\Pi$ are linearly ordered by $\Pi$-ness.
3. In applying $\mathcal{M}$, $S$ has decided that her classification will have the following two characteristics:
   (a) the classification will consist in ruling out some but not all of the possibilities in $P_\Pi$;
   (b) if some possibility $p$ is ruled out, then every possibility that is less $\Pi$ than $p$ will also be ruled out;
4. For $p \in P_\Pi$, any reason $S$ can find for retaining $p$ on the basis of $\mathcal{M}$ is, to roughly the same extent, a reason for retaining $p$'s successor.

(I shall say that a semi-principled classification-method displays tolerance *simpliciter* if it displays tolerance with respect to some parameter or other.)

Under the right circumstances—e.g. case 3 from section 1.1—my handing you the photograph of the bald man is tolerant with respect to the number of hairs on people's heads.

## 2 The main thesis

I am now in a position to state the main thesis of the paper:

> *Main Thesis*
> A representation is vague just in case using it to classify open possibilities consists in deploying a semi-principled classification-method displaying tolerance.

Two remarks are in order. First, the Main Thesis is meant as a substantial claim about the nature of vagueness, not as an analysis of the concept of vagueness. Second, the Main Thesis is meant to apply only to representations that can be used to classify open possibilities. Here are some examples of the sorts of classifications I have in mind:

1. *Photographs*
   I am wondering whether Grampa Lionel was bald. In the context of my inquiry, it is an open possibility that Grampa Lionel was bald and an open possibility that he was not. I use a photograph of Grampa Lionel—who is a clear case of non-baldness—to classify the open possibilities by ruling-out the former.

2. *Drawings*
   I am wondering whether the Blue Nile and the White Nile join in Sudan. In the context of my inquiry, it is an open possibility that they do and an open possibility that they don't. I use my map of Africa to classify the open possibilities by ruling-out the latter.

3. *Utterances*
   I am wondering who won the race. In the context of my inquiry, it is an open possibility that Alice won the race, an open possibility that Bertha won the race and an open possibility that Claudia won the race. I use your utterance of 'Bertha won the race' to classify the open possibilities by ruling-out the first and the third.

4. *Thoughts*
   I am wondering what day of the week it is. In the context of my inquiry, it is an open possibility that today is a Monday, an open possibility that today is a Tuesday and so forth. I suddenly remember that it's a Tuesday. I use my belief that today is a Tuesday to classify the open possibilities by ruling-out all but the second.

Not every representation is used to classify open possibilities. Even though the star representing Khartoum on my map of Sudan is part of what allows me to use the map to classify open possibilities, the star cannot be used to classify open possibilities when considered in isolation. Similarly, even though the name 'Bertha' in your utterance of 'Bertha won the race' is part of what allows me to use the utterance to classify open possibilities, the name cannot be used to classify open possibilities when considered in isolation. I shall call representations that can be used to classify open possibilities *primary* and representations that cannot themselves be used to classify open possibilities but are components of representations that can *secondary*. The Main Thesis might be supplemented with the claim that a secondary representation is vague just in case it is responsible for the vagueness of some primary representation.

The purpose of the remainder of the paper is to explain how the Main Thesis plays out in certain special cases. More specifically, I propose to do the following three things. First, I will argue that the way in which we actually use vague representations to classify open possibilities supports the Main Thesis. Second, I will propose a way of doing semantics for vague representations that is consistent with the Main Thesis. Third, I will argue that the Main Thesis can be used to supply an answer to the Sorites Paradox. The special cases I will consider are representation in public language and mental representation. Extending the proposal to other types of representation should be straightforward.

# 3 Representation in Public Language

It is important to note that the Main Thesis is *not* a claim about meaning. Without supplementation, the Main Thesis is compatible, for instance, with the view that utterances of 'Harry is bald' express perfectly precise propositions. What the Main Thesis implies is that speakers are unable to *use* utterances of 'Harry is bald' to carry out rule-based classifications of the open possibilities, whether or not the proposition expressed by the relevant utterance is, in some sense, rule-based.

The present section is divided into six subsections. In the first I shall try to motivate the claim that representations in public language give rise to semi-principled classification-methods rather than rule-based classification-methods. In the next three I develop a semantics for vague linguistic representations. In the final two subsections I address two of the puzzles that tend to be associated with linguistic vagueness.

## 3.1 Linguistic Competence

There are two good reasons for being uncomfortable with the view that our use of linguistic representation consists in deploying rule-based classification-methods. The first is that reflection on our actual linguistic abilities suggests that utterances involving 'bald' do not, in fact, allow speakers to carry out rule-based classifications, at least not in general. The point emerges most clearly when the circumstances are unusual. You are asked to help identify the murderer in a police line-up. The line-up contains 2001 men. For each $n$ between 0 and 2000, man $n$ has $100n$ hairs on his head (suitably distributed). Hairiness aside, the men are very much a like. All you have to go on is the last words of a dying witness: 'the murderer is bald'. How can you use the utterance to help the police catch their man? The details of the answer will depend on the details of what your linguistic competence consists in. But the outcome can be expected to be this: you will be unable to extract from the witness's utterance a rule that would yield a classification of the open possibilities—that the murderer is man 0, that the murderer is man 1, and so forth. The problem is that there is a certain kind of incommensurability between the information you have at your disposal and the task at hand. The best you can do to classify the open possibilities is make a *decision* about how to use the information at your disposal. You will be forced to balance-off considerations of various sorts. You will try to exclude from the line-up some of the men who are not rendered salient by the utterance and refrain from excluding some of the men who are. You will do your best to strike a balance between excluding enough men that the police finds your participation helpful and retaining enough men that you don't risk clearing the guilty party. In the end, your decision about where to draw the line will involve an arbitrary choice. And your classification method will display *tolerance*, since any reason the utterance might give you to excuse man $k$ is, to roughly the same extent, a reason for excluding his immediate neighbors.

Using the witness's utterance to classify open possibilities will be much easier in less extraordinary circumstances. If the police has narrowed down the suspects to two men, one of which is a clear case of baldness and the other a clear case of non-baldness, your decision about which of the relevant possibilities to exclude on the basis of the witness's utterance will be straightforward.

The second reason for discomfort with the view that our use of linguistic representation consists in deploying rule-based classification-methods arises from the observation that when it comes to sentences involving ‘bald’, it is unclear how the process of linguistic acquisition is supposed to give speakers the ability to classify open possibilities in accordance with any particular rule-based classification-method. For the process must be somehow informed by the speaker’s environment—and, in particular, by the practices of the relevant linguistic community. But even if it is somehow true the speaker’s environment helps determine a particular rule-bases classification-method for, e.g. ‘Harry is bald’, it is unclear that the speaker herself would have the ability to extract the appropriate classification-method from her environment in a fully principled way.

There are, of course, less-than-fully-principled ways for a speaker to settle on a rule-based classification-method. The most flat-footed ways of spelling out this idea are problematic. Were one to claim, for example, that each person’s use of ‘Harry is bald’ is based, once and for all, on a sensibly but arbitrarily chosen classification-rule (e.g. ‘exclude all and only possibilities whereby Harry has more than 35,412 hairs on his head’), one would get all manner of unpleasant results. But more sophisticated versions of the proposal will do better. When particular rules for classifying the open possibilities are chosen on a case by case basis, doing one’s best to strike a balance amongst competing considerations and doing one’s best to make arbitrary choices only when they are least likely to interfere with the prevailing communicative goals, cases of miscommunication will be minimized. Such a proposal would differ only in name from the view that our use of linguistic representation is based on semi-principled classification-methods.

## 3.2 U-truth

No claims about meaning have been made so far. I have defended the thesis that our use of linguistic representation is semi-principled, but this thesis is compatible with various ways of thinking about the meanings of our words. As a prelude to the discussion of semantic issues of the following two subsections, I will characterize a truth-like notion that is based on language-use rather than meaning. I shall call it *use-truth* (or *u-truth*, for short).

> An assertion is *u-true* as used by $S$ just in case the following two conditions are met: ($i$) one of the open possibilities $S$ considers actually obtains, and ($ii$) when the open possibilities are classified by $S$ on the basis of the assertion, the actual possibility is not ruled out. (If only condition ($i$) is met, I shall say that the assertion is u-false as used by $S$.)

(Note that the difference between truth and u-truth is *not* like the difference between truth and truth-according-to-Jones; a better approximation is the difference between true-in-English and true-in-Spanish.)

More often than not, one can expect there to be slight differences between the open possibilities considered by the speaker and the open possibilities considered by her audience. There may even be differences in the ways speaker and audience classify the possibilities. In well-run conversations, the discrepancies will be small enough not to interfere with conversational goals. But even then it might happen that an assertion is u-true as used by the speaker and u-false as used by her audience. In what follows I shall ignore speaker-audience discrepancies for ease of exposition, and speak of $u$-truth and $u$-falsity *simpliciter*. Nothing of substance will turn on this assumption.

Although u-truth often coincides with truth, it is instructive to note that the notions can come apart when the linguistic understanding of conversational participants is in some way limited. Here are some examples:

1. We both believe that 'arthritis' is a disease of the bone rather than a disease of the joints. I have a persistent pain in my thigh but my joints are in perfect health. I say to you 'I have arthritis'. In the right sort of context, my assertion will count as u-true even though one would ordinarily take it to be false.

2. I say to you 'The village barber shaves all and only those who don't shave themselves'. Neither of us notices the lurking contradiction. We fail to consider the question of who shaves the barber: the possibilities we regard as open concern only the various ways in which the non-barbers might be shaved. In the right sort of context, my assertion will count as u-true even though one would ordinarily take it to be false.

A different kind of case in which u-truth and truth might come apart involves counterfactual situations. (Making the point will require a detour which is not essential to the main thread of the paper; uninterested readers may skip ahead to section 3.3.) I begin by introducing some additional notation:

> An assertion is *u-true* relative to world $w$ just in case the following two conditions are met: $(i)$ one of the open possibilities under consideration is compatible with $w$, and $(ii)$ when the open possibilities are classified on the basis of the assertion, the possibility compatible with $w$ is not ruled out. (If only condition $(i)$ is met, the assertion is u-false relative to $w$.)
>
> The proposition u-expressed by an assertion is the set of worlds $w$ such that the assertion is u-true relative to $w$.
>
> The proposition expressed by an assertion is the set of worlds $w$ such that one would ordinarily say that the assertion is true in $w$.

The discrepancy I have in mind arises because there is a systematic kind of mismatch between the proposition u-expressed by an assertion and the proposition expressed by the assertion. Consider an example:

> We are unsure whether the first heavenly body to be visible in the evenings is the last heavenly body to be visible in the mornings. We take two possibilities to be open for the purposes of our conversation: $p_1$ and $p_2$. According to $p_1$ there are two different planets; one of them is the first heavenly body to be visible in the evenings and is referred to as 'Hesperus'; the other is the last heavenly body to be visible in the mornings and is referred to as 'Phosphorus'.

> According to $p_2$ a single planet is the first heavenly body to be visible in the evenings and the last heavenly body to be visible in the mornings, and is referred to as both 'Hesperus' and 'Phosphorus'. I say 'Hesperus is Phosphorus' to you. We classify the possibilities on the basis of my assertion by ruling out $p_1$ and retaining $p_2$.

Whereas the proposition u-expressed by the assertion in this example excludes worlds compatible with $p_1$, the proposition expressed by the assertion in my example is the set of all worlds (since 'Hesperus is Phosphorus' is a necessary truth). Notice, on the other hand, that—worlds incompatible with $p_1$ and $p_2$ aside—the proposition u-expressed by the assertion is precisely the *diagonal* proposition expressed by the assertion. (Roughly, the diagonal proposition expressed by an assertion of $\phi$ is the proposition that would be expressed in a similar context by an assertion of $\ulcorner \phi$ is true$\urcorner$.)

The lesson of this example is that one should think of the proposition u-expressed by an assertion as correlated not with the proposition expressed but with the corresponding diagonal.

## 3.3 Semantics

The purpose of this subsection is to suggest a way of thinking about the meanings of vague expressions. What is special about the view I wish to defend is not the use of non-standard semantic theories. What is special is the way I propose to understand the ascription of a semantic theory to a linguistic community.

A semantic theory, as I shall understand it here, is a compositional assignment of semantic values to sentences. The only type of semantic theory I shall consider is a possible worlds semantics, but nothing of substance hinges on this choice. On a possible worlds semantics, the semantic value of a sentence is a set of possible worlds and the semantic value of a sentential constituent is whatever it needs to be to ensure that the semantic value of a sentence is determined by the semantic values of its parts. Here is one way of spelling out the details:[2]

> Let the intension of a name be a function assigning to each world an object in that world, and let the intension of a one-place predicate be a function assigning to each world a set of objects in that world. The semantic value of a name is taken to be a function from contexts of utterance to name-intensions and the semantic value of a one-place predicate is taken to be a function from contexts of utterance to one-place-predicate-intensions. The semantic value of sentences of the form $\ulcorner P(a) \urcorner$ is then said to be determined by the following rule:
>
> $$v_c^{\mathcal{T}}(\ulcorner P(a) \urcorner) = \{w : v_c^{\mathcal{T}}(a)(w) \in v_c^{\mathcal{T}}(P)(w)\}$$
>
> where $v_c^{\mathcal{T}}(\phi)$ is the result of applying context of utterance $c$ to the semantic value of $\phi$ according to semantic theory $\mathcal{T}$. Finally, a sentence $\phi$ is said to be

[2] I have oversimplified some of the details for ease of exposition. See Lewis (1970) for the full story.

> *true* relative to context of utterance $c$ just in case the actual world is a member of $v_c(\phi)$.

What we have so far is a compositional semantics for a simple language consisting of sentences of the form $\ulcorner P(a) \urcorner$. But the proposal can be developed along similar lines for more complex languages. (A canonical text is Lewis (1970).)

Under what circumstances is it appropriate to ascribe a semantic theory to a given community of speakers? The answer I propose to adopt is based on the notion of fit. Say that a semantic theory $\mathcal{T}$ *fits* an assertion $\alpha$ of $\phi$ just in case the following condition is met:

> [Fit]
> If $\alpha$ is u-true, then $\phi$ is true according to $\mathcal{T}$ (relative to the context of $\alpha$). If $\alpha$ u-false, then $\phi$ is not true according to $\mathcal{T}$ (relative to the context of $\alpha$).

The simplest way of measuring overall fit for $\mathcal{T}$ is by dividing the number of assertions satisfying [Fit] by the total number of u-true or u-false assertions to date. But one could, if one liked, make use of more sophisticated measures. The details are unimportant for present purposes.

> (*First parenthetical remark.* [Fit] needs to be supplemented in various ways. The fit of a semantic theory should not just depend on whether it can be integrated with a theory of assertion as far as truth and falsity are concerned: it should be integrated with a theory of assertion more generally. And not just a theory of assertion: it should be integrated with a general theory of speech acts—including, for instance, a theory of supposition. A first step in this direction would be to broaden one's scope from u-truth and u-falsity to u-truth and u-falsity relative to a possible world. It is tempting to do so by supplementing [Fit] with a condition such as the following: if $\alpha$ is $u$-true relative to $w$, then $w \in v^{\mathcal{T}}_{c(\alpha)}(\phi)$, where $c(\alpha)$ is the context of $\alpha$; and similarly for u-falsity. Unfortunately, it is a consequence of the observation that the proposition u-expressed by an assertion is correlated not with the proposition expressed but with the corresponding diagonal that this proposal won't do. Rather than assessing the fit of a semantic theory, one should assess the fit of a function $\mathcal{F}$ from worlds to semantic theories, where $\mathcal{F}(w) = \mathcal{T}$ is thought of as stating that *as used in w* the language has the semantic properties described by $\mathcal{T}$. One can then say that $\mathcal{F}$ fits an assertion $\alpha$ of $\phi$ just in case the following condition is met: if $\alpha$ is u-true relative to $w$, then $w \in v^{\mathcal{F}(w)}_{c(\alpha)}(\phi)$, where $c(\alpha)$ is the context of $\alpha$; and similarly for u-falsity. Since $\mathcal{F}$ is part semantics and part metasemantics the result is that one ends up assessing the fit of one's semantics and metasemantics as a package. And even this is not enough as a full characterization of fit, since it doesn't take account of pragmatic phenomena such as conversational implicature. Accordingly, one must assess the fit of one's semantics, metasemantics and pragmatics as a package. All of this will be ignored in what follows for the sake of simplicity.)

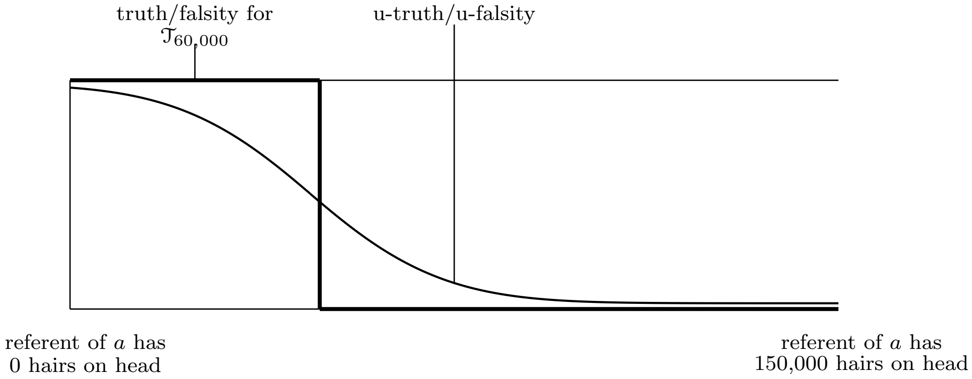

Figure 1: Comparison between relative frequency of u-truth to u-falsity and relative frequency of truth to falsity according to $\mathcal{T}_{60,000}$.

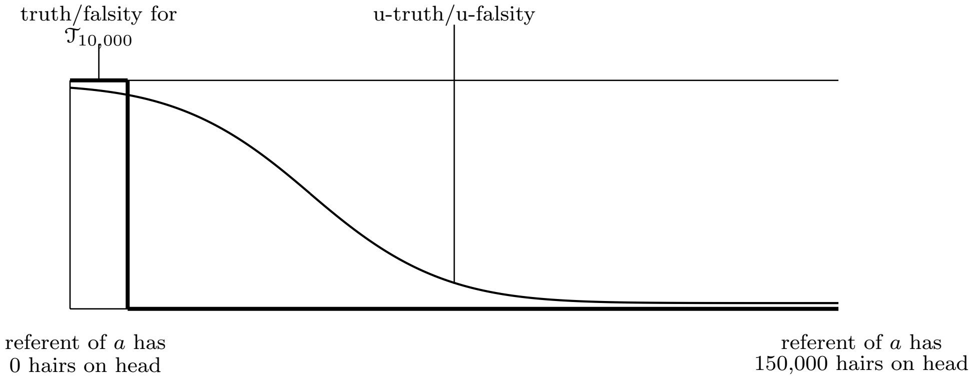

Figure 2: Comparison between relative frequency of u-truth to u-falsity and relative frequency of truth to falsity according to $\mathcal{T}_{10,000}$.

Here is an illustration of the condition on fit. Let $\mathcal{T}_n$ be a semantic theory according to which a sentence of the form $\ulcorner a \text{ is bald}\urcorner$ is true just in case the referent of $a$ is a person with $n$ hairs on her head or less. The smooth curves in figures 1 and 2 represent the relative frequency of u-truth to u-falsity of assertions of sentences of the form $\ulcorner a \text{ is bald}\urcorner$ (in contexts of utterance meeting some given constraint). The angled curves represent the relative frequency of truth to falsity of sentences of the form $\ulcorner a \text{ is bald}\urcorner$ according to $\mathcal{T}_{60,000}$ and $\mathcal{T}_{10,000}$ (relative to contexts of utterance meeting the given constraint). The lesson that emerges is that even though neither of the two semantic theories delivers perfect fit, $\mathcal{T}_{60,000}$ fits the facts about linguistic usage somewhat better than $\mathcal{T}_{10,000}$, at least with respect to the range of cases considered.

By individuating contexts finely enough and postulating significant variations in semantic value from context to context, it is possible to construct a semantic theory that delivers perfect fit—but at a considerable cost in simplicity, since the resulting theory will be thoroughly unsystematic. It is also possible to construct a semantic theory of unrivaled simplicity (e.g. by making every sentence true relative to every world and context)—but at a considerable cost in fit. A good semantic theory will strike a good balance between simplicity and fit. The better the balance, the better the theory.

> (*Second parenthetical remark.* It is useful to see how the proposal plays out with respect to certain interesting cases. [Rigidity] Semantic theories according to which 'Hesperus' is a rigid designator can be expected to deliver a better combination of simplicity and fit than their rivals because they will fit assertions of counterfactuals (e.g. 'Had Hesperus had a different orbit, it might not have been the first heavenly body to be visible in the evenings') with no corresponding loss in simplicity. [Twin Earth Cases] Semantic theories according to which 'water' refers to $H_2O$ as used by Earthlings and to XYZ as used by Twearthlings can be expected to deliver a better combination of simplicity and fit than their rivals because they will better fit assertions of, e.g. 'There's water in this jug' and 'This may not be water even though it behaves just like water' with no corresponding loss in simplicity. [Compositionality] The village barber example in section 3.2 is a case in which a semantic theory could only fit the relevant assertion at a considerable cost in simplicity; one can therefore expect 'the village barber shaves all and only those who don't shave themselves' to be false according to semantic theories delivering the best combinations of simplicity and fit. [Grue and Bleen] Assigning expressions gruesome semantic values comes at a cost in simplicity. [Divisions of Linguistic Labor] There are two kinds of cases to consider. The first occurs when experts make finer distinctions than non-experts but the usage of non-experts is compatible with the usage of experts; here semantic theories making the finer distinctions can be expected to deliver better overall fit with no corresponding loss in simplicity because they better fit the assertions of experts without failing to fit the assertions of non-experts. The second kind of case occurs when the usage of non-experts is sometimes incompatible with the usage of experts, as in the arthritis case in section 3.2. If the instances of discrepancy are isolated enough, then considerations of simplicity will suffice to favor semantic theories accommodating the expert usage. But when instances of discrepancy are sufficiently widespread additional resources are needed to secure the result that the usage of experts is to be favored. One way to proceed is by assessing the combined merits of one's semantics and metasemantics, as in the first parenthetical remark. There is then room for arguing that combined theories that sacrifice fit with respect to non-expert assertions but are able to motivate the relevant semantic assignments in terms of, e.g. an (imperfect) referential chain from the experts to the non-experts should be preferred over combined theories that deliver better fit with respect to non-expert assertions but are unable to motivate the semantic assignments in terms of a suitable referential chain.)

It is tempting to suppose that the semantic theory delivering the best combination of simplicity and fit should be ascribed to the relevant linguistic community with no further ado. I would like to suggest that this is a temptation that ought to be resisted.

A first reason to resist is that the facts about linguistic usage make it unreasonable to expect that there will be a non-arbitrary way of identifying a semantic theory that delivers the 'best' combination of simplicity and fit. Figures 1 and 2 can be used to illustrate the point. In light of the shape of the curve representing the ratio of u-truth to u-falsity, the differences in fit between very similar semantic theories can be expected to be extremely small—small enough to be unstable. It could well turn out, for example, that $\mathcal{T}_n$'s degree of fit is just higher than $\mathcal{T}_{n+1}$'s, but take the relevant linguistic community to have a slightly different membership, or count a slightly different set of events as assertions, or use a slightly different measure of overall fitness and the situation is reversed. And considerations of simplicity are unlikely to help. Even if one is prepared to grant, for instance, that a semantic theory can be simpler than its rivals by assigning semantic values that come closer to 'carving nature at the joints', it seems implausible to suppose that any one way of classifying people into bald and non-bald comes closer to carving nature at the joints than all the rest. If anything, considerations of simplicity can be expected to make matters worse, since different semantic theories might be favored by ways of balancing off considerations of simplicity and fit that one could only decide amongst by making arbitrary choices.

A second reason for resisting the temptation is that one can expect any reasonably simple semantic theory not to deliver a very good fit. One way to see this is by noticing how poorly the curve corresponding to $\mathcal{T}_{60{,}000}$ in figure 1 approximates the curve corresponding to u-truth and u-falsity. In order to get a better approximation, one would have to complicate $\mathcal{T}_{60{,}000}$ by building complex patterns of contextual variation into the semantic value of 'bald'. The result is that semantic theories delivering good combinations of simplicity and fit are unlikely to yield very accurate descriptions of the relevant linguistic practice.

The lesson, it seems to me, is that one should change the way one thinks about the ascription of a semantic theory to a linguistic community. Rather than thinking that there is a single semantic theory containing the whole semantic truth about the language in question, one should describe the practice of speakers by providing information about the degree of fit of one or more semantic theories.

Which semantic theories need be considered and how much information needs to be provided about their degree of fit depends entirely on the goals of one's inquiry. For instance, by considering a single theory and saying of its degree of fit that it is about as good as it gets amongst reasonably simple theories, one can provide substantial information about the linguistic practice of the relevant community in a relatively succinct way. By considering more than one semantic theory and describing their degrees of fit in great detail one can increase the amount of information one provides about linguistic practice, but at a cost in succinctness. What is important is that one be clear about the sorts of claims that are licensed by the attribution of a degree of fit to a semantic theory. Conspicuously, one should resist the inference from the claim that $\mathcal{T}$ is at least as good a semantic theory as any of its rivals to the claim that speakers will generally use assertions to classify the open possibilities in accordance with the truth-conditions supplied by $\mathcal{T}$.

## 3.4 Truth

An immediate consequence of the view I have been defending is that it makes no sense to say of a sentence that it is true *simpliciter* (even when the context of utterance has been fixed). The most one can say is that the sentence is true or false according to a given semantic theory and that the semantic theory delivers such and such a combination of simplicity of strength. Fortunately, it doesn't follow that the notion of truth *simpliciter* must be given up. I will argue in this section that one can say of an *assertion* that it is true or false *simpliciter*.

Before saying what it is for an assertion to be true *simpliciter*, I must say what it is for a sentence to be true relative to a context of assessment. Following John MacFarlane and others,[3] I distinguish between a sentence's context of utterance and its context of assessment. The context of utterance is what fills the contextual parameters in semantic components of sentences: it is, for example, what determines the referent of 'I' in my utterance of 'I am hungry'. Accordingly, it is the context of utterance that fills the contextual parameter in the notion of truth-according-to-a-semantic-theory. The context of assessment, on the other hand, is a set of mutually exclusive open possibilities with respect to which the truth of a given sentence is to be assessed once the sentence's contextual parameters have been filled in by a context of utterance.

The notion of truth relative to a context of assessment can be characterized as follows. Sentence $\phi$ is true relative to context of assessment $\mathcal{A}$ just in case the following three conditions are satisfied:

1. One of the possibilities in $\mathcal{A}$ actually obtains.

2. A unique classification $\mathcal{C}$ of the possibilities in $\mathcal{A}$ is rendered salient by semantic theories with reasonable aptness.

3. The actual possibility is not ruled out by $\mathcal{C}$.

(If the first two conditions are satisfied but third is not, $\phi$ is *false* relative to $\mathcal{A}$; if the first condition fails to be satisfied, $\phi$ *carries false presuppositions* relative to $\mathcal{A}$; if the second fails to be satisfied, $\phi$ is *defective* relative to $\mathcal{A}$.)

What do I mean by aptness? A semantic theory is *apt* to the extent that it delivers a good combination of simplicity and fit. What do I mean by salience? Classification $\mathcal{C}$ of $\mathcal{A}$ is rendered *salient* by semantic theories of reasonable aptness if $\mathcal{C}$ does significantly better than any of its rivals in satisfying the following two conditions:

1. *Quantity* An sizable majority of reasonably apt semantic theories induce classifications that agree with $\mathcal{C}$.

2. *Quality* Semantic theories inducing classifications that disagree with $\mathcal{C}$ are dominated in aptness by semantic theories in $S$ inducing classifications that agree with $\mathcal{C}$.

[3] See MacFarlane (2005) and Egan et al. (2004).

The classification *induced* by a semantic theory $\mathcal{T}$ is the classification that consists in retaining all and only possibilities in $\mathcal{A}$ that are compatible with some world in $v_c^{\mathcal{T}}(\phi)$, where $\phi$ is the sentence under consideration and $c$ is the context of utterance.

> (*Third parenthetical remark.* The notion of truth relative to $\mathcal{A}$ can be generalized to a notion of truth-in-$w$ relative to a context $\mathcal{A}$, where $w$ is a possible world. Say that a sentence $\phi$ is true in $w$ relative to $\mathcal{A}$ just in case the following three conditions are satisfied: (*i*) one of the possibilities in $\mathcal{A}$ is compatible with $w$; (*ii*) a unique classification $\mathcal{C}$ of the possibilities in $\mathcal{A}$ is rendered salient by semantic theories with reasonable aptness; and (*iii*) the possibility compatible with $w$ is not ruled out by $\mathcal{C}$. Accordingly, one can say that the *proposition* expressed by $\phi$ relative to $\mathcal{A}$ is the set of worlds $w$ such that $\phi$ is true in $w$ relative to $\mathcal{A}$. Truth is closed under classical consequence, in the following sense: if $\psi$ is a classical consequence of $\phi$ and $\phi$ is true relative to $\mathcal{A}$, then $\psi$ is true relative to $\mathcal{A}^{\phi}$, where $\mathcal{A}^{\phi}$ is the proposition expressed by $\phi$ relative to $\mathcal{A}$. Relative to contexts of assessment with respect to which $\phi$ and $\psi$ are not vague—i.e. contexts of assessment with respect to which assertions of $\phi$ and $\psi$ will never be defective—one gets the following additional result: if $\psi$ is a classical consequence of $\phi$ and $\phi$ is true relative to $\mathcal{A}$, then $\psi$ is true relative to $\mathcal{A}$. It is important to bear in mind, however, that these results depend on the decision to use standard possible-worlds semantics as our semantic theories. Different choices of semantic theories might deliver different results.)

Some examples will make clear how the various definitions are supposed to work. In each case I will consider the question of whether 'Harry is bald' is true relative to context of assessment $\mathcal{A}$.

- *Case 1* $\mathcal{A}$ consists of two possibilities, one of which is actual. The first is a possibility whereby Harry is a clear case of baldness; the second is the possibility whereby Harry is a clear case of non-baldness.

  Although reasonably apt semantic theories will often disagree about the extension of 'bald', the vast majority of them can be expected to agree in inducing classification $\mathcal{C}$, which consists of ruling out the second of the two possibilities in $\mathcal{A}$ and keeping the first. Moreover, if any reasonably apt semantic theory disagrees with $\mathcal{C}$ it will be strictly dominated in aptness by a semantic theory that agrees with $\mathcal{C}$. This suffices to render $\mathcal{C}$ salient. Accordingly, 'Harry is bald' is true relative to $\mathcal{A}$ if the first possibility is actual, and false relative to $\mathcal{A}$ if the second possibility is actual.

- *Case 2* $\mathcal{A}$ consists of a large number of possibilities, one of which is actual. The possibilities can be divided into two sharply demarcated categories: according to every possibility in the first category, Harry has few enough hairs on his head to count as a clear case of baldness; according to every possibility in the second category, Harry has enough hair to count as a clear case of non-baldness.

Exactly analogous to case 1. ‘Harry is bald’ is true relative to $\mathcal{A}$ if a possibility in the first category is actual, and false relative to $\mathcal{A}$ if a possibility in the second category is actual.

- *Case 3* $\mathcal{A}$ consists of two possibilities, one of which is actual. The first is a possibility whereby Harry is a clear case of baldness; the second is a possibility whereby Harry is a borderline case of baldness. The context of utterance is that of an assertion of ‘Harry is bald’ in a situation such that: ($a$) speaker and audience can both see that there are two men sitting at the table next to theirs; ($b$) one of the men is a clear case of baldness and the other is a borderline case of baldness; and ($c$) it is common knowledge that Harry is one of the men and that only the speaker knows which.

  Most reasonably apt semantic theories can be expected to fall into one of two categories. Those in the stricter category will induce classification $\mathcal{C}^s$, which consists in keeping the first possibility and ruling out the second. Those in the laxer category will induce classification $\mathcal{C}^l$, which consists in keeping all possibilities. If any reasonably apt semantic theory falls outside one of these categories it can be expected to be dominated in aptness by a semantic theory that doesn’t. There is no difference in salience between $\mathcal{C}^s$ and $\mathcal{C}^l$ so far, but the context of utterance will make the difference. For semantic theories according to which the extension of ‘bald’ depends on whether some comparison class is salient in the context of utterance will achieve a substantial increase in fit at a comparatively low cost in simplicity, and amongst semantic theories that are sensitive to salient comparison classes, those that set a high standard for baldness in contexts such as the one under discussion can be expected to dominate in fit. This is sufficient to render $\mathcal{C}^s$ salient. Accordingly, ‘Harry is bald’ is true relative to $\mathcal{A}$ if the first possibility is actual, and false relative to $\mathcal{A}$ if the second possibility is actual.

  (The complementary situation, in which Harry is a borderline case of baldness according to the first possibility and a clear case of non-baldness according to the second, yields a complementary result: ‘Harry is bald’ is true relative to $\mathcal{A}$ if the first possibility is actual, and false relative to $\mathcal{A}$ if the second possibility is actual.)

- *Case 4* $\mathcal{A}$ consists of a large number of possibilities, one of which is actual. For any amount of hair a normal person might have, $\mathcal{A}$ contains the possibility that Harry has that amount of hair.

  Any two reasonably apt semantic theories disagreeing about the extension of ‘bald’ will induce a different classification of the possibilities. As a result, no single classification of the possibilities is rendered salient. So ‘Harry is bald’ is defective relative to $\mathcal{A}$.

Now that I have explained what it is for a sentence to be true relative to a context of assessment, I am in a position to say what it is for an assertion to be true *simpliciter*: an assertion is true just in case the sentence asserted is true relative to the possibilities that are regarded as open by conversational participants for the purposes of the assertion.

Analogously, an assertion is false or defective or carries false presuppositions just in case the sentence asserted is false or defective or carries false presuppositions relative to the possibilities that are regarded as open by conversational participants for the purposes of the assertion. (As in the case of u-truth, there might be slight differences in the possibilities that different conversational participants regard as open for the purposes of the assertion. In well-run conversations, the differences will be small enough not to interfere with the conversational goals. As before, I will ignore such discrepancies for expositional purposes.)

The notion of truth *simpliciter* is the result of bringing together a semantic notion and a pragmatic notion. The semantic notion is the notion truth relative to a context of assessment; the pragmatic notion is the notion of a possibility's being regarded as open for the purposes of an assertion. Although the semantic notion hasn't received much study, the pragmatic notion has been studied at great length in the past three decades. (See, for instance, the articles collected in Stalnaker (1999).) In particular, there are good reasons for thinking that the possibilities regarded as open for the purposes of an assertion are subject to a *Rule of Accommodation*:

> The possibilities regarded as open for the purposes of an assertion should evolve so as to make it the case that the assertions that conversational participants accept in the course of the conversation turn out to be true. (All of this, subject to constraints spelled out in Lewis (1979).)

Thus, the Rule of Accommodation demands that if I assert "Fred's children are asleep" and my audience accepts the assertion, the possibilities which are regarded as open for the purposes of my assertion must include only possibilities in which Fred has children, thereby assuring that my assertion carries no false presuppositions.

The Rule of Accommodation will play an important role in keeping assertions of 'Harry is bald' from being defective. Consider, for instance, an assertion of 'Harry is bald' in a context in which nothing special is presupposed about the amount of hair on Harry's head. Suppose, in particular, that for any amount of hair a normal man might have on his head, conversational participants take the possibility that Harry has that amount of hair to be open at the time in which the assertion takes place. The lesson of Case 4 above is that the assertion will be defective if the possibilities regarded as open for the purposes of the assertion coincide with the possibilities which conversational participants take to be open at the time in which the assertion takes place. But the Rule of Accommodation will keep this from happening. It will demand that the possibilities which are regarded as open for the purposes of the assertion evolve so as to make it the case that the assertion is non-defective. The most natural way for this to happen is for conversational participants to treat Harry's not being a borderline case of baldness as a presupposition of the assertion, and hence take some of the possibilities whereby Harry is a borderline case of baldness *not* to open for the purposes of the assertion. The situation will then be as illustrated as Case 2. And on the assumption that the actual possibility hasn't been presupposed away, the assertion will be either true or false, depending on the amount of hair on Harry's head. (Just which possibilities are presupposed-away is something it is up to each conversational participant to decide, so different conversational participants may take different possibilities to be open for the purposes of the assertion. But as long as the differences are small enough not to affect common goals, there will be no risk of miscommunication and the conversation will carry on without incident.)

The characterization of truth I have provided in this section is, of course, vague. Conspicuously, there is vagueness in the notion of reasonable aptness, vagueness in the notion of salience and vagueness in what count as possibilities which are open for the purposes of the assertion. But the vagueness is harmless, for the following two reasons. First, the proposal's verdicts about truth-status are largely stable, in the sense that making the characterization more precise won't usually affect the truth-status assigned to ordinary assertions. Second, the vagueness in question does not interfere with the proposal's ability to elucidate the nature of the transition between cases in which a vague representation applies and cases in which it doesn't. It does not, for example, interfere with the proposal's ability to deliver the following explanation: although there is broad agreement amongst the most reasonable semantic theories that ⌜a man with $n$ hairs on his head is bald⌝ is true when $n = 0$, the agreement gradually becomes weaker as $n$ increases; eventually, there will be no agreement at all; but as $n$ continues to increase there will be more and more agreement about the sentence's falsity, and by the time $n$ reaches 150,000 there will be broad agreement amongst the most reasonable semantic about the sentence's falsity.

It is worth noting that the present proposal is compatible with a version of the view that linguistic understanding consists of tacit semantic knowledge. For although one couldn't claim that linguistic understanding consists of tacit knowledge of some particular semantic theory, one could claim that it consists of tacit knowledge about the fit of one or more semantic theories. Fortunately, this is an issue one can afford to remain neutral about for present purposes.

## 3.5 The Sorites Paradox

Here is an instance of the Sorites Paradox. (I ignore matters of hair distribution for the sake of simplicity.)

> It is tempting to think that each of the following is true:
>
> S1 A man with 0 hairs on his head is bald.
>
> S2 For any $n$ between 0 and 1999, if a man with $100n$ hairs on his head is bald, a man with $100(n + 1)$ hairs on his head is also bald.
>
> S3 A man with 200,000 hairs on his head is not bald.
>
> But S1–S3 jointly entail a contradiction.

How is the contradiction to be averted? From the present perspective, the basic answer is straightforward: S2 is false relative to any context of assessment.[4] But the Sorites hasn't

[4] To see this, note that when the contextual parameter is fixed in the intended way—e.g. no unusual comparison classes, no relevant partitions of men made salient by context—the following conditions obtain:

been fully addressed until the temptation to accept S2 in the first place has been accounted for.

An explanation emerges from the observation that in ordinary contexts, sentences involving 'bald' give rise to classification-methods that display *tolerance* with respect to the number of hairs on people's heads (in the sense of tolerance characterized in section 1.3). Consider an assertion of 'Harry is bald' and assume we prevent the Rule of Accommodation from coming into effect by insisting that, for any $n$ between 0 and 200,000, it to be an open possibility that Harry has $n$ hairs on his head. One is, of course, unable to extract from the assertion any particular classification of the possibilities. The best one can do is classify the possibilities in a way which is arbitrary but sensible in light of the considerations at hand. And, crucially, tolerance will be displayed: any reason the assertion gives one for not ruling out a possibility whereby Harry has $k$ hairs on his head is, to roughly the same extent, a reason for not ruling out a possibility whereby Harry has $k + 100$ hairs on his head. This makes it natural to think that something like the following must be true:

> If man with $k$ hairs on his head is bald, then so is a man with $k + 100$ hairs on his head.

For when one focuses one's attention on a possibility whereby Harry has $k$ hairs on his head and reflects on the reasons the assertion gives one for not ruling out this possibility, one finds that those same reasons apply to a possibility whereby Harry has $k + 100$ hairs on his head, to roughly the same extent. Note, for comparison, that it is *not* natural to suppose that the following must be true in general:

> If man with $k$ hairs on his head is bald, then so is a man with $k + 40,000$ hairs on his head.

For when one focuses one's attention on a possibility whereby Harry has $k$ hairs on his head and reflects on the reasons the assertion gives one for not ruling out this possibility, one does *not* generally find that those same reasons apply to a possibility whereby Harry has $k + 40,000$ hairs on his head, to any comparable extent.

I would like to suggest that it is this sort of reasoning that explains the temptation to think that S2 is true. It is worth noting that my explanation is based on a story about the ability of speakers to use linguistic representations, not on a story about the semantic properties of linguistic representations. This is as it should be. For it is the linguistic competence of speakers that informs their intuitions. Whether or not semantics comes into the picture will depend on the relationship between semantics and linguistic competence. There is, however, a semantic observation to be made about the case at hand: any sentence of the form 'A man with $k$ hairs on his head is bald but a man with $k + 1$ is

(*i*) S2 is necessarily false according to the vast majority of reasonably apt semantic theories, (*ii*) each of the few dissenters is strongly dominated in aptness by a large number of non-dissenters, and (*iii*) each of the few dissenters takes S2 to be necessarily true. So the classification that consists of ruling out all open possibilities is rendered salient.

not' is either false or defective relative to any context of assessment.[5] There is therefore an interesting respect in which it is impossible to produce a counterexample to S2.

## 3.6 Omniscient Speakers

You and Julia and are sitting in a room with 2001 men before you. For each $n$ between 0 and 2000, man $n$ has $100n$ hairs on his head. Julia is an omniscient speaker: she knows everything there is to know about the matter at hand. She is also fully attentive and fully cooperative, and wishes to be as relevantly informative as she can. For each $n$, you ask her whether the sentence ⌜man $n$ is bald⌝ is true relative to a fixed context of utterance and a fixed context of assessment. Both contexts are reaonable. You ask her to answer 'yes' or 'no'. What will her answers be?

If the proposal I have been defending is along the right lines, Julia will be unable to use her mighty knowledge to come up with a list of answers to your questions without making some arbitrary choices along the way. The best she can do is make a decision about how to answer that is sensible in light of the situation at hand. She will make sure not to answer 'yes' to the $n + 1$th question unless she has answered 'yes' to the $n$th question. She will make sure she answers 'yes' to the first few questions and 'no' to the last few questions. She will take into account all the information there is to be had about the usage of 'bald' in your linguistic comunity. But, in the end, her decision about where to draw the line will involve an arbitrary choice.

Sensing that Julia may have delivered her answers under duress, you decide to ask your questions again, but this time allowing her to give any answer she deems appropriate, or say nothing at all. Her response to your first question is as follows:

> It would be a mistake to suppose that there is a *rule* to be extracted from your question—a rule that could be used to classify 'yes' or 'no' answers into those that are 'correct' (or 'permissible' or 'mandated') and those that aren't. Rather than giving rise to a rule, your question gives rise to a semi-principled classification method because there is a certain kind of incommensurability between the information I have at my disposal and the task at hand. The best I could do to come up with 'yes' or 'no' answers to your questions is make a *decision* about how to use the information at my disposal. In the case of the question you have just asked—Is 'man 0 is bald' true?—facts about the linguistic practice of your community make the answer 'yes' salient. In the case of the question you intend to ask last—Is 'man 2000 is bald' true?—facts about the linguistic practice of your community make the answer 'no'

[5] To see this, note that when the contextual parameter is fixed in the intended way, the following conditions obtain: (*i*) the potential counterexample is necessarily false according to the vast majority of reasonably apt semantic theories, (*ii*) for each of the few dissenters there are a large number of non-dissenters of at least comparable aptness, and (*iii*) each of the few dissenters takes the potential counterexample to be necessarily true; so if some unique classification of the open possibilities in the context of assessment is rendered salient by the classifications induced by reasonably apt semantic theories, it will be the classification that consists of ruling out all open possibilities.

> salient. But bear in mind that, because of the vagueness of your term ‘true’, the classification-method I employ displays tolerance. In particular, any reason I have for thinking that any particular ‘yes’ or ‘no’ answer has been rendered salient in the case of your $n$th question will be, to roughly the same extent, a reason for thinking that that same answer has been rendered salient in the case of your $n+1$th question. This will have the effect that any decision on my part to give differing ‘yes’ or ‘no’ answers to any particular pair of consecutive questions will involve an arbitrary choice. And the differing answers are likely to be misleading, since they might tempt you to think that there is something special about the relevant pair of questions.

## 4 Mental Representation

As a result of meeting Harry at a party, you acquire the belief that Harry is bald. Let HB be the mental representation that is thereby placed in your belief-box. I would like to suggest that when it comes to the task of classifying open possibilities, HB gives rise to semi-principled classification method rather than a rule-based classification-method. (Note that this is not a claim about HB’s semantic properties—it is not, for example, a claim about HB’s truth-conditions—it is a claim about how you are able to use HB to classify open possibilities.)

The point emerges most clearly when you are placed in an unusual situation. Harry has been accused of committing a crime, and you are asked to help identify him in a police line-up. The line-up contains 2001 men. For each $n$ between 0 and 2000, man $n$ has $100n$ hairs on his head (suitably distributed). Hairiness aside, each of the 2001 men looks just like Harry. How can you use HB to help the police catch their man? The details of the answer will depend on how relevant information is stored in your cognitive system. (For instance, things will go one way if information is stored in the form of a mental picture of Harry and another if information is stored in the form of a memory in which you say to yourself ‘That guy is bald’.) But the outcome can be expected to be this: you will be unable to extract from HB a rule that would yield a classification of the open possibilities—that Harry is man 0, that Harry is man 1, and so forth. As usual, the problem is that there is a certain kind of incommensurability between the information you have at your disposal and the task at hand. The best you can do to classify the open possibilities is make a *decision* about how to use the information at your disposal. You will be forced to balance-off considerations of various sorts. You will try to exclude from the line-up some of the men who are not rendered salient by HB and refrain from excluding some of the men who are. You will do your best to strike a balance between excluding enough men that the police finds your participation helpful and retaining enough men that you don’t risk clearing the guilty party. In the end, your decision about where to draw the line will involve an arbitrary choice. And your classification method will display *tolerance*, since any reason HB might give you to excuse man $k$ is, to roughly the same extent, a reason for excluding his immediate neighbors.

Using HB to classify the open possibilities would be much easier in less extraordinary circumstances. Suppose one of the men dining at the table next to ours is a clear case of baldness and the other is a clear case of non-baldness. You learn that one of them is Harry and turn towards them in hopes of learning which. Even if the men have very similar features, a decision about how to classify the open possibilities on the basis of HB will be relatively straightforward. As before, the details will depend on how relevant information is stored in your cognitive system. Suppose, for example, that the information is stored in the form of a mental picture. Then the picture will render a unique classification of the open possibilities salient. Since Harry is bald and your mental picture is more or less accurate, HB will render the possibility whereby Harry is the bald man salient. So the classification will consist in ruling out the possibility whereby Harry is the clear case of non-baldness.

According to the Main Thesis, a representation is vague just in case it gives rise to a semi-principled classification-method displaying tolerance. So the upshot of the example is that HB is a vague mental representation. It is important to note that HB is said to be vague in the same sense that representations in public language are said to be vague. This is an instance of the fact that the present proposal treats vagueness as a general phenomenon of representation, rather than a special feature of public language.

The example focuses on a case in which you acquire the belief that Harry is bald as a result of meeting him at a party. But it is worth emphasizing that a similar story would emerge on other ways of coming to believe that Harry is bald. You may come to believe that Harry is bald by making an inference from the observation that all of his male relatives are bald. Or you may come to believe that Harry is bald because of someone's assertion of 'Harry is bald'. In either case, one can expect the relevant mental representation to give rise to a semi-principled representation method displaying tolerance, just as before. (Of course, if you acquired the belief that Harry is bald by learning that there are precisely 200 hairs on his head, the situation will be somewhat different since the information you have at your disposal in this case will *not* give rise to the usual sort of incommensurability when it comes to classifying open possibilities in the police line-up.) Notice, moreover, that the focus on *belief* in the example is inessential. I could have instead considered a case in which HB is placed in your desire-box, or your fear-box.

Nothing has been claimed so far about the semantic properties of vague mental representations. But friends of the Language of Thought hypothesis are welcome to ascribe semantic properties to vague mental representations by following the strategy I recommended in section 3.3 for the case of public language. This is how the story would go, in broad outline:

> Just as in the case of public language, start by characterizing a notion of u-truth that describes the classifications of open possibilities that the agent actually carries out on the basis of her mental representations. Next, develop a theory of syntax for mental representations (on the basis of, e.g. considerations of generativity) and make a choice about the sorts of compositional semantic theories that will be used to describe the semantic properties of mental representations.

> Explain what it is for a semantic theory to *fit* the facts about u-truth just as in the case of public language, and say that a semantic theory is *apt* to the extent that it delivers a good combination of simplicity and fit. Resist the temptation to think that there will be a maximally apt semantic theory that should be ascribed to the agent with no further ado; instead, describe the semantic features of an agent's representations by saying of one or more semantic theories that they are apt to such-and-such degrees. Finally, characterize notions of truth and falsity relative to a set of open possibilities by considering which ways of classifying the open possibilities are rendered salient by reasonably apt semantic theories, just as in the case of public language.

There are, however, two important *caveats*. The first is that in the case of thought there are no pragmatic mechanisms to get us from truth relative to a context of assessment to truth *simpliciter*. Instead, the possibilities with respect to which a given use of a mental representation is assessed are the possibilities the agent sets out to discriminate amongst. The second *caveat* arises as follows. When characterizing the notion of u-truth in the case of public language, we were able to help ourselves to talk of open possibilities with no further ado, but now we need to be more careful. For presumably part of what it is for an agent to regard a possibility as open is for the possibility to be represented by a mental representation suitably located in the agent's cognitive system. So we are caught in a circle: according to the story above, the semantic properties of a mental representation depend on the possibilities it is used to classify, but according to what has just been noted, which possibilities are being classified depends on the semantic properties of mental representations. This means that there is only so much that the story above can be used to achieve. In the case of public language, the analogous story could be used to ascertain the semantic properties of representations in public language by taking for granted the semantic properties of mental representations. In the case at hand, the most we can do is use the story to place constraints on the semantic properties of mental representations. The constraints will be of two kinds. On the one hand, the story constrains the *form* that attributions of semantic properties to mental representations should take: rather than ascribing a semantic theory to an agent with no further ado, one should describe the semantic properties of the agent's representations by saying of one or more semantic theories that they are true to such-and-such degrees. The second type of constraint is a restriction on the claims one is allowed to make about the aptness of a semantic theory. Specifically, one's claims must be such that they yield the following reflective equilibrium: the semantic theories one claims to be apt should, by and large, be the semantic theories that turn out to be apt by the lights of the story above when the open possibilities with respect to which u-truth is defined are determined on the basis of semantic theories one claims to be apt. Although these constraints are non-trivial, they won't give us much information about the fit of any particular semantic theory. This is as it should be, lest we get the result the semantic properties of mental representations are independent of causal connections between the agent and her environment.

If the story above is properly supplemented, one can expect a version of the Sorites

Paradox based on mental representations to allow for the same sort of treatment as the corresponding public language-based Sorites. If, for example, one were to replace each of S1–S3 from section 3.5 with a corresponding sentence of Mentalese, one would be able to explain away the contradiction by observing that, relative to any set of open possibilities, the major premise will be either false or defective. And, just as in the case of a public-language based Sorites, the initial appeal of the major premise would be addressed by the observation that sentences of Mentalese involving the concept BALD give rise to classification-methods that display tolerance relative to the number of hairs on people's heads.

# 5 Concluding Remarks

I hope to have achieved three things. First, I hope to have shown that the ways in which we actually use vague representations to classify open possibilities lend support to the hypothesis that vague representations give rise to semi-principled classification-methods displaying tolerance. Second, I hope to have sketched a workable strategy for talking about the semantic properties of representations giving rise to semi-principled classification-methods. Third, I hope to have shown that the assumption that vague representations give rise to semi-principled classification-methods displaying tolerance can be used to supply an answer to the Sorites Paradox. It seems to me that these three points together make a reasonable case for the Main Thesis.

An account of vagueness based on the Main Thesis has the advantage of overcoming the two limitations I described at the outset. On the one hand, it treats vagueness as a phenomenon of representation in general, of which vagueness in public language is merely a special case. On the other, it characterizes the nature of vague representation without postulating unexplained boundaries and without obstructive vagueness in the metalanguage.

# References

Bräuerle, R., et al., eds. (1970) *Semantics from Different Points of Vew*, Springer-Verlag.

Dorr, C. (2003) "Vagueness without Ignorance," *Philosophical Perspectives* 17.

Egan, A., J. Hawthorne, and B. Weatherson (2004) "Epistemic Modals in Context." In Preyer and Peter (2004).

Evans, G., and J. McDowell, eds. (1976) *Truth and Meaning*, Clarendon Press, Oxford.

Graff, D. (2000) "Shifting Sands: An Interest-Relative Theory of Vagueness," *Philosophical Topics* 28:1.

Keefe, R., and P. Smith, eds. (1996) *Vagueness: A Reader*, MIT Press, Cambridge, MA.

Laurence, S. (1996) "A Chomskian Alternative to Convention Based Semantics," *Mind* 105, 269–301.

Lewis, D. (1969) *Convention: A Philosophical Study*, Harvard University Press, Cambridge, MA.

Lewis, D. (1970) "General Semantics," *Synthese* 22, 18–67. Reprinted in Lewis (1983).

Lewis, D. (1979) "Score-Keeping in a Language Game," *The Journal of Philosophical Logic* 8, 339–59. Reprinted in Bräuerle et al. (1970); reprinted in Lewis (1983).

Lewis, D. (1983) *Philosophical Papers, Volume I*, Oxford.

MacFarlane, J. (2005) "Making Sense of Relative Truth," *Proceedings of the Aristotelian Society* 105, 321–39.

McGee, V., and B. McLaughlin (1994) "Distinctions without a difference," *The Southern Journal of Philosophy* XXXIII, 203–251.

Preyer, G., and G. Peter, eds. (2004) *Contextualism in Philosophy*, Oxford University Press, Oxford.

Sainsbury, R. M. (1990) "Concepts without Boundaries." Inagural Lecture, given at King's College London. Reprinted in Keefe and Smith (1996).

Stalnaker, R. C. (1979) "Assertion," *Syntax and Semantics* 9, 315–322. Reprinted in Stalnaker (1999).

Stalnaker, R. C. (1999) *Context and Content*, Oxford University Press, Oxford.

Williamson, T. (1994) *Vagueness*, Routledge, London and New York.

Wright, C. (1976) "Language-Mastery and the Sorites Paradox." in Evans and McDowell (1976); reprinted in Keefe and Smith (1996).