Radical Anti-Disquotationalism

Andrew Bacon*

February 6, 2019

Abstract

A number of 'no-proposition' approaches to the liar paradox find themselves implicitly committed to a moderate disquotational principle: the principle that if an utterance of the sentence 'P' says anything at all, it says that P (with suitable restrictions).[1] I show that this principle alone is responsible for the revenge paradoxes that plague this view. I instead propose a view in which there are several closely related language-world relations playing the 'semantic expressing' role, none of which is more central to semantic theorizing than any other. I use this thesis about language and the negative result about disquotation to motivate the view that people do say things with utterances of paradoxical sentences, although they do not say the proposition you'd always expect, as articulated with a disquotational principle.

Consider a self-referential utterance, u, of the sentence 'u is not true'. According to one widespread and appealing intuition when one makes a semantically paradoxical utterance such as u one simply does not succeed in saying anything. Call this the *no proposition theory*.

No proposition theorists reject the disquotational assumption that utterances of 'u is not true' say that u is not true on the grounds that some utterances of 'u is not true' do not say anything at all. However, they may nonetheless subscribe to a qualified version of the principle that says that if an utterance of 'u is not true' says *anything at all* it says that u is not true. In this paper I shall show that such views are in the unfortunate position of not being able to state their own view. An examination of these paradoxes suggests a view in which paradoxical utterances such as u *do* say things, although they do not say what you might expect them to (in this case, that u is not true). I shall show, moreover, that this phenomena falls out of general considerations about the relation between language and the world, and that as a result failures of disquotational principles are much more widespread than many have thought.

In this paper we will be primarily be investigating the corner of philosophical space that accepts classical logic, understood broadly to include the classical rules for quantification into sentence position.[2]

*I'd like to thank John Hawthorne, Bernhard Salow, Gabriel Uzquiano and Bruno Whittle for their comments on various earlier versions of this paper. I'd also like to thank Cian Dorr for spotting several errors in the appendix.

[1] Restrictions might exclude sentences which express different propositions in different contexts, such as sentences involving indexical expressions.

[2] Of course, neither of these assumptions are uncontentious; but see Prior [33] and Williamson [45] for a defense of the intelligibility of quantification into sentence position, and Williamson [46] for general methodological remarks about the application of classical logic to the liar paradox. For a recent approach

1 No Proposition Accounts of the Paradoxes

The most straightforward version of the no proposition theory maintains that sentences like 'u is not true' are completely meaningless, and thus that any attempt to say something by making an utterance of this sentence would fail. Call this the *non-contextual* view. On the non-contextual view, then, all utterances of a paradoxical sentence are equally bad.

However, a central component of many extant no-proposition accounts of the semantic paradoxes is the contention that while one utterance of a sentence might fail to say anything, a distinct utterance of the *very same* sentence might result in something being said — perhaps even something true and important to the theorist.[3] Suppose u is an utterance of the sentence 'u is not true'. If u doesn't say anything because it is paradoxical then, presumably, u is neither true nor false. So, in particular, u is not true. The important thing to note is that I made this last point — that u is not true — by producing a different token of the sentence 'u is not true'. When *I* produce a token of 'u is not true' I say something true and theoretically enlightening.

This is how Williamson describes the process which brought about the change between the utterance u and my utterance:

> "We start with one set of correlative meanings for 'say', 'true' and 'false'; we use them to construct a sentence that says nothing in that sense of 'say'; but reflection on that sentence causes normal speakers to give 'say', 'true' and 'false' a new set of correlative meanings, much like the previous ones except that the sentence in question says something in the new sense of 'say'; the process can be repeated indefinitely." [43] p15

According to Williamson consideration of the semantic paradoxes forces us to recognise a more expansive use of the word 'says' (and by extension 'true' and 'false'). Self-referential utterances involving these more expansive uses of these words force us to recognise even more expansive uses, and this process repeats indefinitely. As I find Williamson's way of setting up the issues congenial, I will use his theory as a springboard for my investigation into 'no proposition' accounts of the liar. Germs of this idea trace back at least as far as medieval scholars and has a number of contemporary adherents.[4]

Since it is granted that some utterances of 'u is not true' do say something, there is a *prima facie* onus on us to explain why u in particular doesn't say anything. The explanation cannot simply be that u is self-referential, for there are self-referential utterances that do say things; an email beginning 'I'm just sending you this message because...' surely says something even though it's self-referential.

An explanation for why u doesn't say anything is available if we help ourselves to the following principles:

Utterance Truth If an utterance says that P then it is true if and only if P

to the liar that relaxes classical propositional logic, see Priest [31], Brady [8], Field [16]. For a discussion of approaches that relax the classical rules for the propositional quantifiers see Bacon, Hawthorne and Uzquiano [6].

[3] See, for example, Williamson [43], Goldstein ([20]) and Gaifman [18] among others.

[4] A non-exhaustive list includes Bar-Hillel [7], Whiteley [41], Prior [32], Goldstein ([20]), Gaifman [18], Simmons [35], Weir [40], Rosenkrantz and Sarkohi [34]. Note that Williamson does not quite fit the traditional picture: he argues that the paradoxes produce a shift in the conventional meaning of the word 'says'. Thus both the content and character associated with a sentence can vary from utterance to utterance. The important point, for my purposes, is the variability of the content of tokens of the same sentence.

Moderate Disquotation If an utterance of 'u is not true' says anything at all, it says that u is not true.

Given our assumption that u is an utterance of 'u is not true', these two principles entail that u does not say anything. For if u did say something then according to the second principle, since u is an utterance of the sentence 'u is not true', it would say that u was not true. Then by the first principle it would follow that u was true if and only if it wasn't. This is, of course, all perfectly consistent with the possibility that other utterances of 'u is not true' do say things; if v is a distinct utterance of 'u is not true' then all we can derive from the assumption that it says something is that v is true if and only if u isn't true.

The first principle is explicitly endorsed by Williamson.[5] Typically we speak of people, not utterances, as saying things; however it is a fairly harmless abuse of convention to extend the notion to the utterances that facilitated the saying. A couple of points need to be stressed about this convention, however. Firstly, speech reports are notoriously flexible — I can use the expression 'so and so said that P' to cover things they merely intended to communicate, or simply things they hinted at. I am primarily interested in the slightly more regimented uses of 'says' that is the focus of much semantic theorizing.[6]

Secondly, in the unregimented sense we typically count people as saying obvious logical consequences of things that they say. In our more regimented sense we should understand 'what u was used to say' rather strictly as something like 'the strongest thing u was used to say', or else we should not expect the first principle to hold (see the discussion in Andjelkovic and Williamson [1] p230-232).[7]

Utterance Truth is related to a slogan of truth conditional semantics: that what a sentence means in a context are the conditions under which it is true. However, in reality it encodes little more than a choice to use the words 'says' and 'true' in a related way when applying them to utterances: that an utterance is true if whatever it says is true. If we permit ourselves a device that allows us to quantify into sentence position, as Williamson does, one can give a definition of utterance truth from utterance saying as follows:

> u is true if and only if: (i) for some P such that u says that P, P, and (ii) for no P such that u say that P, is it not the case that P.[8]

A bit more colloquially, u is true iff it says something true and nothing false. Note that the colloquial way of paraphrasing this definition makes use of the notions of propositional

[5] See principle T on p. 12. The difference is that Williamson talks about sentences in contexts rather than utterances. He notes that this principle is importantly different from the T-schema; indeed Utterance Truth is consistent in classical logic.

[6] See, for example, Grice [21]. Even with all the clarifications and stipulations that philosophers perform before using the word, its important to note that it still retains a substantial amount of flexibility.

[7] If there isn't a strongest proposition that u says, one may simply take the conjunction of the propositions u says instead. Given the assumption that the propositions form a complete Boolean algebra, this always determines a unique proposition although. Those who prefer to theorise in terms of structured propositions, and who thus deny the Boolean assumption, will typically countenance many such conjunctions. Even if there are multiple such conjunctions, they will all be materially equivalent (indeed, logically equivalent) since they only differ over the order of the conjuncts. The paradoxes we will be considering pose just as much a problem for the view that each sentence strictly says several materially equivalent propositions as for the view that each sentence strictly says one thing (cf Prior's Theorem in section 2.2).

[8] This definition may have to be modified if we want to make it friendly to propositional temporalism: the view that there are propositions that are sometimes true and sometimes false. The natural modification would be: u is true $=_{Def}$ for some P, u said that P, and P was true at the time u was uttered and for any P, if u said that P then at the time u was uttered it was the case that P. I will not make any attempt to control for this complication in what follows.

truth and falsity, even though they make no appearance in the above definition: here and elsewhere I shall often use such paraphrases for convenience, but it is important to note that truth and falsity, as they apply to propositions, make no appearance in the sentences they are paraphrasing.

If we uniformly substitute 'true' with our definition of utterance truth ('says something true and nothing false') in Utterance Truth, we get the plausible claim that if an utterance says that P then it says something true and nothing false if and only if P. Thus we can prove Utterance Truth from our definition and the assumption that no utterance says more than one thing. Indeed, this assumption is already guaranteed by our convention to understand 'says' so that it relates u to P only if P is he conjunction of everything u said in the unregimented sense.

In combination with the claim that u said that u is not true, Utterance Truth implies that u is true if and only if it isn't. Since Utterance Truth plays a role in the derivation of a contradiction, one might be tempted to lay the blame on Utterance Truth. This would be a mistake. One might object to Utterance Truth on the grounds that it involves an incorrect account of the relation between truth and saying. But one cannot object that Utterance Truth is inconsistent, or the source of the paradox, since it is evidently just a consequence of my choosing to define utterance truth in terms of utterance saying.

The second principle, Moderate Disquotation, was also needed to explain why u doesn't say anything. As a simple model to demonstrate that no inconsistency follows without it, note that if u had said that snow was white, then u would have said something, and it would be straightforwardly true assuming our definition of utterance truth. Without this moderate form of disquotation there is no logical guarantee that u doesn't say anything. Thus, without Moderate Disquotation, the no-proposition view does not follow from considerations of the paradoxical utterances like u, as one might have expected.

It should be stressed that Moderate Disquotation is not a schema, it is just a principle about a particular sentence. Some instances of the more general schema are hardly plausible: if you utter the sentence 'I'm hungry' you've said something, but you haven't said that *I'm* hungry you've said that *you* are hungry. This moderate disquotational principle implies that the no-proposition theory is not a familiar form of contextualism, exhibited by sentences involving indexicals like 'I' and ordinary context sensitive words like 'tall'. Two utterances of the sentence 'Fred is tall', for example, can be used to say different things. In the case of 'u is not true' two utterances of this sentence can differ with regard to whether they express a proposition at all but given the moderate form of disquotation above no two utterances of 'u is not true' that express a proposition can express different propositions: if they both express propositions, they must both express the proposition that u is not true.[9]

A sentence is context sensitive in the more familiar sense when it can be used to say different things in different contexts. Contextualism in this sense has also been applied to the liar paradox, but I will not classify these views as 'no-proposition' theories.[10] Such views will typically maintain that 'u said something' is false in some contexts but *true* in others,

[9]Note that some contextualists (e.g. Simmons [36]) often seem to speak as if u does express a proposition; u is defective, rather, because the proposition it expresses is neither true nor false. If the proposition itself was defective in some way, then, given the point that all utterances of 'u is not true' that say anything say the same thing, all of the other utterances expressing that proposition would be defective. This would be disastrous since the contextual no-proposition theorist wants to make utterances of the sentence 'utterance u is not true' themselves when they describe their view.

[10]Examples of the standard form contextualism includes Parsons [30], Burge [9], and Glanzberg [19]. Such views will not be the focus of this paper, although they face similar problems to the one I'll raising for: they have trouble expressing their own view (see Williamson [44] section V)

much like contextualists about singular quantification will maintain that the sentence 'there is beer' is false in some contexts but true in others. It would be a mistake to conflate the standard contextualist with the no-proposition theorist just as it would be a mistake to conflate the nihilist about beer with a contextualist about quantification over beer.

In this paper I will endeavour to show that, despite appearances, u actually *does* say something, even if it cannot be the thing you'd articulate using an disquotational principle.[11]

My main goal in section 3 will be to show that Moderate Disquotation is untenable: it straightforwardly gives rise to further paradoxes. This undermines our original argument for thinking that u does not say anything. I argue instead for an alternative thoroughly anti-disquotational view. We begin with the observation that verbs like 'says that' are vague, semantically plastic and context sensitive. These considerations naturally lead us to the thesis of Semantic Pluralism: that there are a large number of similar but equally natural language-world relations, with none uniquely playing the honorific role of the 'expressing relation' that takes a central place in semantic theorizing. The view is radical in the following sense: once we have noted that there are several distinct semantically important relations I shall argue that few of these relations relate even utterances of the (apparently unproblematic) sentence 'snow is white' to the proposition that snow is white.

The rest of the paper will be structured as follows. In section 2, I briefly outline a consistent framework for regimenting discussions about propositions, and outline an important theorem in this system concerning what can be said, due to Arthur Prior. In section 3 I argue that any theory endorsing certain instances of moderate disquotation are untenable. I go on to motivate the thesis I called Semantic Pluralism in section 4, and explain how it is unfriendly to disquotationalist principles across the board. Semantic Pluralism is brought to bear on the semantic paradoxes in section 5. In the appendices, a formal theory of compositional expression is outlined and some of the ideas of the paper are formalized and shown to be consistent.

[11] A tentative (but suggestive) argument for this conclusion is already available: the above theory provides an attractive account of contingent liars that would otherwise pose a puzzle for no-proposition accounts. If I am standing in an unknown room and I see the sentence 'the only sentence token written on the board in room 101 is untrue' written on an otherwise blank board, I appear to be in a position to rule out the epistemic possibility that the board in room 101 has nothing written on it but the sentence 'snow is white'. This is a quite general fact about the reaction of English speakers to seeing that string of symbols written on a board, assuming they take it to be a reliable source of truths. The default explanation of this fact would be that the sentence token expresses a proposition inconsistent with this possibility. Indeed one might think that it is facts like these that constitute what it means for a sentence to express a proposition that's inconsistent with that kind of possibility. The theory I sketched corroborates the default explanation whereas the no-proposition theory does not, for if you are in fact standing in room 101 the token would not have expressed a proposition. Even if this is not a decisive reason to think that this sentence token says something, even when you are standing in room 101, there is clearly *some* relation between that token and a proposition, and one would want some story about analogous paradoxes for that relation. For example, you can still talk about the set of worlds that competent English speakers will rule out upon accepting the token as trustworthy, or the proposition the token *would* have expressed if I had wheeled the board into the hallway (so it was no longer located in room 101). For a discussion of the latter case, see Zardini [47] p563-4. Analogous paradoxes involving these relations suggest that simply denying that a proposition is expressed won't get to the heart of the matter.

2 Paradoxes of Indirect Discourse

2.1 Propositions and quantification into sentence position

The semantic paradoxes concern notions, such as truth and saying, that relate language to the world. We'll draw a sharp line between the semantic paradoxes and what Ramsey has termed the *logical paradoxes* which, to put it glibly, concern the world alone. These include paradoxes of propositions and properties that do not rely on the semantic principles connecting language to the world.[12] If there were a proposition, p, identical to the proposition that p is not true, for instance, then we can generate a non-semantic version of the liar paradox resembling the semantic one in derivational form. Yet it's plain that we will not find a diagnosis of this paradox by studying semantic connections between utterances and propositions.

If we are to tackle the semantic paradoxes without getting tangled up with the logical paradoxes, it is clear that we need a more concrete and precise framework for theorizing about propositions. Some recent metaphysicians have advocated for the idea that certain murky questions about propositions and properties may be replaced by more perspicuous questions formulated by employing quantifiers that can directly bind variables taking sentence and predicate position, as we did in the last section.[13] This system is called *higher-order logic*; a formal presentation of it can be found in the appendix. In this setting, the analogue of the propositional truth predicate is an *operator*, T. But its existence is trivial: it is just another truth functional connective like negation (its truth table is the trivial one). In a standard system of higher-order logic it may be defined by identifying T with the λ expression $\lambda X\, X$, which denotes the operator that maps every proposition to itself.[14] It is then a simple consequence of the usual principles governing λs that $TP = P$, from which an analogue of the T-schema follows: $TP \leftrightarrow P$.[15] The aforementioned paradox involving propositions becomes a theorem to the effect that no proposition is identical to the proposition that it is not true.[16]

$$\neg\exists P(P = \neg TP)$$

But it's also evident that this notion of propositional truth is entirely eliminable: any appearance of TP in a sentence can be faithfully replaced by just P, because standard systems of higher-order logic prove that $TP = P$.

Since higher-order logic is consistent, we have a clear framework in which to formulate questions about propositions. It must be noted, however, that further principles concerning the granularity of propositions maybe added to higher-order logic that render it inconsistent. For example, principles stating that propositions are structured, and thus effectively

[12]Ramsey understood 'logic' broadly enough to include set theory, and thus counted the set theoretic paradoxes as logical. I'll be focusing on puzzles that are purely logical even by todays standards.

[13]A recent and forceful instance of this is attitude is manifest in Williamson [45]. See also Prior [33].

[14]There are several axiomatic treatments of higher-order logic that build in various theses about propositional granularity (usually the thesis that there are only two propositions). In the following I will work in the extremely minimal axiomatic system of higher-order logic H described in Bacon [3]. This system contains the laws of classical propositional logic, the usual axioms and rules for the first-order quantifiers, and their analogues for quantification at other types, and laws that ensure that λ expressions behave appropriately.

[15]Note here that $=$ as used above is a sentential connective: it should be thought of as standing to sentences as the usual identity relation stands to names, and is subject to the obvious analogue of Leibniz's law. It can be seen that $TP \leftrightarrow P$ follows from $TP = P$ and $P \leftrightarrow P$ by applying Leibniz's law to the left-hand-side of the latter sentence.

[16]This theorem is proved by noting that, given Leibniz's law, $P = \neg TP$ and $P \leftrightarrow P$ implies $P \leftrightarrow \neg TP$. We may then note that this is inconsistent with $TP \leftrightarrow P$.

as fine-grained as sentences of a public language, unsurprisingly allow one to construct self-referential propositions as one might construct self-referential sentences, e.g., via diagonalization.[17] The moral must be different here, however: the structure principles are inconsistent on their own in higher-order logic, and the T-schema analogue is innocent as it does not appear in the derivation of this inconsistency.[18] So much the worse for structured theories of propositions. Alternative doctrines about propositional granularity such as *Booleanism* — the view that Boolean equivalents are identical — are straightforwardly consistent in higher-order logic.

One might wonder about the fate of first-order theories of propositions. We have been content to paraphrase sentences formalized using quantification into sentence position informally by quantifying singularly over propositions, employing a propositional truth predicate where necessary: it would be unfortunate if this practice was inconsistent.

Let's formalize the first-order a theory of propositions a little more explicitly. In higher-order logic there are types for sentences (type t), singular terms (type e), and functional types like $e \to t$ and $t \to e$, which encode expressions that take singular terms as arguments to produce sentences in the former case, and sentences to produce singular terms in the latter. A retraction pair from t to e, is a pair F and G of type $e \to t$ and $t \to e$ respectively, which compose to form the identity operation on type t: $\lambda P\, F(G(P)) = \lambda P\, P$. The existence of a retraction is not provable in an entirely minimal theory of higher-order logic, but it's a natural extension of the idea that there aren't more things of type t than of type e (since any injection from type t to type e ought to have a left-inverse).[19] A first-order theory of propositions can be formulated with a primitive predicate *true* of type $e \to t$ that states that a given individual is a proposition and is moreover true, and a primitive operation *that* of type $t \to e$ that produces a name for a proposition when applied to a sentence (*that* P may be pronounced 'the proposition that P'). We say that an individual x is a proposition if, for some P, $x = \text{that}\, P$. The propositional T-schema informally says that the proposition that P is true if and only if P: $\text{true}(\text{that}(P)) \leftrightarrow P$. This follows from a slightly stronger claim — $\text{true}(\text{that}(P)) = P$ — which, for instance, captures the thought that saying that the proposition that snow is white is true should be the same as just saying that snow is white. This is concisely captured in higher-order logic by by the claim that *true* and *that* form a retraction pair from t to e:

First-order Propositions $\lambda P\, \text{true}(\text{that}(P)) = \lambda P\, P$.

Thus the first-order theory of propositions is interpretable by any retraction from t to e.[20] It follows that the first-order theory of propositions is consistent, as the consistency of higher-

[17] Specifically, the principle Dorr [12] calls *Structure* is inconsistent in higher-order logic: $\forall XY \forall xy(Xx = Yy \to (x = y \land X = Y))$ where X and Y range over operators (type $t \to t$) and x and y over propositions (type t). Whittle [42] has also shown how first-order theories of structured propositions are subject to diagonalization worries. These diagonalization arguments can be seen as a different take on the Russell-Myhill paradox (see, e.g., Hodes [26]), which are often stated in terms of cardinality.

[18] As we noted, the analogue of the T-schema, $TP \leftrightarrow T$, is just a theorem of higher-order logic when T is defined by $\lambda X\, X$.

[19] Natural axiomatic extensions of H do not prove the existence of a left-inverse of every injection (see the appendix of Bacon [3]), so it is a proper extension of the idea, and not a merely consequence. But it is a natural one, given that every injective set-theoretic function has a left inverse.

[20] It's worth noting that, for this reason, it's consistent with first-order theory of propositions that Julius Caesar is a proposition. But this may be seen as a benefit of the theory: one does not need to postulate a special class of abstract entities to play the role of propositions, the theory works just as well with any retraction pair, allowing any individual to play the role of a proposition relative to some choice of retraction.

order logic with a retraction from t to e is easily shown to be consistent.[21] The above observation can also be used to provide, relative to a choice of retraction, a translation between the fragment of higher-order logic involving quantification into sentence position and a first-order theory of propositions.[22] This will prove useful since English doesn't contain quantifiers that take sentence position, and paraphrasing sentences involving quantification into sentence position with singular quantification over propositions greatly improves readability. The first-order theory of propositions similarly contains a theorem to the effect that there are no self-referential propositions:

$$\neg\exists x(x = \text{that}\neg\text{true}(x))$$

That is, nothing is identical to the proposition that it isn't true.[23]

2.2 Prior's Theorem

I have been theorizing with the familiar notion of what a person has said, as captured by indirect speech reports of the form 'so-and-so said that P' (understood in the slightly regimented way I outlined earlier). From this we can indirectly introduce the notion of what an utterance was used to say — roughly, if a person said that P, and this was facilitated in some essential way by an utterance or sentence token of some other kind, then that utterance can be thought of as 'saying that' P. Once this is understood we can then introduce utterance truth in the way described earlier using quantification into sentence position.

My strategy in the following will be to shed light on the semantic paradoxes by figuring out what people have said when they make utterances of paradoxical sentences. There are, of course, variant paradoxes that are not directly treated by this approach — paradoxes involving a more technical piece of philosophical vocabulary: the notion of a sentence 'semantically expressing' a proposition in a language and context, and the correlated notion of truth in a language and context. According to the simplest theory of this relation a sentence semantically expresses the proposition that P in a language and context if and only if an utterance of that sentence made in that language and context would result in one saying that P. Given the simplest theory, the focus on figuring out what people have said is well placed. One might object that the informal notion of what is said by a person is too flimsy to do the theoretical jobs that we need the semantic expressing relation for. Nonetheless, I shall adopt the simplest theory as a placeholder in what follows, as I think it is a good enough approximation of the truth. If we can figure out how the semantic paradoxes behave on this theory we are on track to figuring out how they work more generally.

[21] Any full model of higher-order logic, in which the cardinality of the domain of type t is no greater than the cardinality of the domain of type e establishes this.

[22] This involves systematically replacing variables of type t with variables of type e, quantifiers over type t with a restricted quantification of type e, restricted by the property of being a proposition. The following provides a translation of the propositionally quantified fragment of higher-order logic with an operator S (described in the next section), into a first-order fragment of higher-order logic with an operator S. First we associated each variable of type t, p_i, with a variable of type e, x_i, and the translation, $\phi \mapsto \phi^*$, proceeds as follows $(p_i)^* = \text{true}(x_i)$, $(\phi \wedge \psi)^* = \phi^* \wedge \psi^*$, $(\forall p_i \phi)^* = \forall x_i(Prop(x_i) \to \phi^*)$, $(\neg\phi)^* = \neg(\phi^*)$, $(S\phi)^* = S(\phi^*)$. (More sophisticated translations are available that also allow for a comparable theory of operators and connectives. For example, there will be an operation of individuals of type $e \to e$ that represents any given operator, so that, for instance, $\neg$ becomes $\lambda x\, \text{that}\neg\text{true}x$.)

[23] Proof sketch: Suppose that $x = \text{that}\neg\text{true}(x)$. We may thus apply truth to both sides to conclude that $\text{true}(x) = \text{true}(\text{that}(\neg\text{true}(x)))$. Finally, since $\lambda P\, \text{true}(\text{that}(P))$ is the identity, $\lambda P\, P$, way may conclude that $\text{true}(x) = \neg\text{true}(x))$, which entails the obviously inconsistent $\text{true}(x) \leftrightarrow \neg\text{true}(x)$.

Since indirect speech reports will play an important role in this investigation we will begin by looking at an important limitative theorem concerning their logic due to Arthur Prior (Prior [32] and Prior [33]). It is important in the following to keep in mind the difference between indirect and direct speech reports. 'Alice said that snow is white' is an example of the former — it is formalized using an operator expression, and it states that Alice bears a relation to a certain proposition. 'Alice said 'snow is white' ' is an example of the latter, and it merely says that Alice uttered 'snow is white'. Prior's theorem does not, without extra assumptions, imply anything about what sentences can or cannot be uttered.

Prior's theorem is formulated against the background of a fragment of higher order logic, known as 'quantified propositional logic', a language containing the propositional connectives that also allows quantification directly into the position occupied by a sentence.[24] Formally speaking quantified propositional logic adds to the syntax of propositional logic a countable set of propositional variables, p_i, and a special quantifier $\forall$. Each propositional variable is a well formed formula, and in addition to the standard clauses for building up complex well formed formulae we stipulate that if p_i is a propositional variable and ϕ a well formed formula, $\forall p_i \phi$ is well formed. Finally we will be interested in the extension of this language by a unary connective, S, which prefixed to any well formed formula, ϕ, produces a well formed formula $S\phi$. S here will stand in for 'Alice said at t that'. Within this context Prior proves a general theorem concerning arbitrary operators which results in seemingly paradoxical results when taken to be about the particular operator S.

Within this system he derives the following theorem:

Prior's Theorem $S\forall p(Sp \to \neg p) \to \exists p(Sp \land p) \land \exists p(Sp \land \neg p)$

The proof, in Polish notation, can be found in [32].

The premises required to derive Prior's theorem are surprisingly minimal.[25] An intuitively complete axiomatization of quantified propositional logic would require a richer system, however to derive Prior's theorem one only needs, in addition to the standard principles of classical propositional logic, the principle of universal instantiation for the propositional quantifiers which allows you to infer '...ϕ...' from '$\forall p$ ...p...'.[26] We can axiomatise this system as follows:

CL All substitution instances of classical tautologies and the rule of modus ponens.

PUI $\forall p\phi \to \phi[\psi/p]$ where ψ is substitutable for p in ϕ.[27]

Following Prior, we can paraphrase the reasoning behind the theorem in ordinary English using singular quantification over propositions. Suppose that Alice says at t that everything that Alice says at t is untrue. It follows that she said something true: if she hadn't, then everything she said at t was untrue – but that's exactly what she said, so something she said is true after all. But if something she said at t was true then she was simply not speaking truthfully when she said that everything she said at t was untrue – that is to say, she said something untrue too, namely, that everything she said at t is untrue.

[24]This is sometimes misleadingly called 'propositional quantification', although the use of these quantifiers does not commit you to the existence of propositions which are singular entities and can which can be quantified over using the ordinary first order quantifiers.

[25]Prior doesn't explicitly note that the necessary system is this weak, although it is immediate from inspection of his proof.

[26]In what follows we assume $\forall$ and $\to$ as primitive, and treat $\exists$ and $\land$ as defined (note that $\bot$ may be defined as $\forall p p$, and negation in terms of $\to$ and $\bot$).

[27]Intuitively 'ψ is substitutable for p in ϕ' just means that no free sentential variables in ψ get bound when substituted into ϕ for p.

Two points are worth reiterating at this juncture. (i) Prior's paraphrase involves singular quantification into the argument of a predicate such as in '... is not the case', presumably committing us to some sort of first-order theory of propositions. Prior's theorem, on the other hand, does not involve quantification over proposition-like entities, it involves quantification into the position that a sentence occupies which is entirely compatible with there being no abstract objects. (ii) Prior's paraphrase makes reference to something resembling a propositional truth predicate (i.e. '... is the case'), and falsity predicate ('... is not the case') whereas Prior's Theorem employs a device that quantifies directly into sentence position and therefore makes no use of propositional truth or falsity. Despite these differences, I will adopt these paraphrases as well — to do otherwise would be cumbersome and would involve taking liberties with English, and, as we have noted above, they are harmless relative to the existence of a retraction.

As Prior himself emphasizes, the theorem is very general. Because no assumptions about S were made, one can interpret S as any number of attitudes such as 'thinks that', 'fears that', 'means that', 'writes that', 'hopes that', 'desires that' and so on. To illustrate how surprising some of these instances are, let me highlight an interpretation of S that delivers particularly puzzling instances of Prior's theorem: 'it sounded as though Alice said, at t, that ...'. Plugging this into Prior's theorem, and paraphrasing in English, gives us:

- If it sounded at t as though Alice said that everything it sounded as though Alice said at t is false, then there's a truth and a falsehood it sounded as though Alice said at t.

We'll see shortly that many responses to original paradox don't have much to say about this variant.

3 Against Moderate Disquotation

Anyone who accepts classical logic and the logic of quantification into sentence position must accept Prior's Theorem. Given straightforward paraphrases of the propositional quantifiers in Prior's Theorem into English using singular quantifiers, however, it seems that we must accept the following apparent falsehoods:

1. Necessarily, if anyone says that everything they've ever said is false then they said something else at some point.

2. Necessarily, if Alice says at t that everything she's said at t is false, then she's said at least two things at t (something false and something true.)

It is worth pausing to reflect on the strength of these results, and contrast them with more familiar forms of the liar paradox. Indeed, the literature on the liar usually focuses on paradoxes formulated using language relative notions, like the notion, mentioned earlier, of a sentence expressing a proposition in a language (possibly relative to a context), or the notion of a sentence being true in a language (relative to a context).

It is important to note how the notion of saying differs from these more technical notions. Saying is something a person does, not a sentence. In archetypal cases of saying, one says something with the help of a public language sentence. But not all cases of saying need be like this. For example, suppose you are attempting to communicate with someone in a foreign country (and you don't speak the language): typically one establishes *ad hoc* conventions in order to say things. Indeed, you can in principle say things with your eyebrows, and other facial expressions, if the context is right. More importantly, even in the more conventional cases of saying involving public language sentences, saying is not language relative. You can say that snow is white using the English sentence ‘snow is white’ or the German sentence ‘schnee ist weiss’. Thus saying, and the corresponding notion of utterance truth are not language relative.

Theorists who focus on paradoxes formulated using the language relative notions will often draw morals and offer consolations like the following:[28]

> The liar paradox shows that the property of being a true sentence of $\mathcal{L}$ (for instance) cannot be expressed by a predicate of $\mathcal{L}$ itself. However, this does not mean that the property of being a truth of $\mathcal{L}$ is inexpressible, for there could be a more expressive language, $\mathcal{L}^+$, which can express the property of being true in L. Of course, $\mathcal{L}^+$ cannot express the property of being true in $\mathcal{L}^+$, but there are yet more expressive languages for that, and so on.

By contrast, Prior’s paradox shows that there is a proposition that Alice cannot say uniquely at t *simpliciter*. But that means that she can’t say it with a sentence of English, or some extension of English, English$^+$, or with her eyebrows, or with interpretive dance, and so on. This conclusions is striking, and the puzzle is not alleviated at all by making speeches like the one above.

3.1 What Alice uttered and what she said

What platitudes do 1 and 2 apparently contradict? One platitude is that Alice could perfectly easily have uttered, at t, the *sentence* ‘everything Alice is saying at t is false’ and have not simultaneously uttered anything else. This much is absolutely clear and any theory must accommodate this fact. What must be contested, then, is whether this fact contradicts 2. All 2 implies is that if, in uttering this sentence, Alice had said, at t, that everything Alice says at t is false, then she would have said something else at t as well. It is crucial to bear in mind the point emphasized earlier that the connection between the sentence one utters and what one says (if anything) by uttering that sentence, is not straightforward and is certainly not governed by a simple disquotational principle: if John said that he was hungry, for example, he almost certainly wouldn’t have achieved this by uttering the sentence ‘he was hungry’, for he would use this sentence only to ascribe past hungriness to someone else. Yet this is exactly what one would get if you unthinkingly applied a naïve disquotational principle.

Prior’s theorem, in combination with this platitude, therefore also places some constraints on this relation that holds between utterances and propositions. Let us suppose that Alice does, in fact, utter the sentence ‘everything I am saying at t is false’ at t, and nothing else and let us ask what she thereby said. Here are the answers to this question that are consistent with Prior’s theorem:

1. Alice didn’t say anything at all when she made this utterance. In particular, she didn’t say that everything Alice says at t is false.[29]

2. Alice did say that everything she says at t is false, but in doing so she also said something else (and moreover, she said at least two propositions with opposite truth

[28] For a representative example, see Eklund [14].

[29] This corresponds to the standard no-proposition line, see the references in section 1 .

value.)[30]

3. Alice said something, perhaps exactly one thing, but whatever it was that she said, it wasn't that everything she said at t is false.[31]

The first upshot, which can be seen just by surveying these options, is that we cannot apply the naïve disquotational method to determine what Alice said in this case. We must reject:

Disquotation Anyone who utters the sentence 'everything Alice says at t is untrue' (with the right intentions, etc) thereby says that everything Alice says at t is untrue and nothing else.

A second striking feature is that according to two of the three options Alice says something completely unexpected. According to the third option Alice says exactly one thing. But since, according to that view, this proposition is not the proposition that everything Alice said at t is false we are left with a mystery as to which proposition she did say. According to the second option Alice says two things, although, at least one of these things is a mysterious proposition. For the time being I shall refer to both of these views as *mysterious proposition* views.

It seems, then, that the no-proposition view (the first option) has a distinct advantage since it seems to avoid these mysterious propositions. In this section I shall argue that this is not the case: the general thought that we don't express an unexpected proposition is untenable. We can state the idea that utterances like Alice's should say the expected proposition, if they say anything at all, more precisely in terms of the following weakened disquotational principle (a variant of the principle Moderate Disquotation):

Moderate Disquotation* Anyone who utters the sentence 'everything Alice says at t is untrue' and says anything at all by it, says that everything Alice says at t is untrue.[32]

To completely rule out mysterious propositions being expressed one would have to add that the proposition that everything Alice says at t is false is the *only* thing one would end up saying with an utterance of this sentence, if you were to say anything at all. However, I shall actually be refuting the weaker thesis encapsulated by the above principle. Note that, in so far as these questions are discussed at all by contemporary no-proposition theorists, this moderate form of disquotation is often implicitly being assumed, not for context sensitive language in general, but for the kind of context sensitivity that arises with sentences like the one Alice uttered.[33]

[30] This seems to be the view of Thomas Bradwardine, and some subsequent medieval scholars. This option is less well represented in the modern literature, although Cian Dorr has taken the multiple-proposition line on the paradoxes in some talks circa 2011, which will be taken up further in a planned monograph.

[31] This option is also not very well represented in the modern literature, but see Nicholas J.J. Smith [37].

[32] MODERATE DISQUOTATION* is clearly motivated by the same general sort of intuition that MODERATE DISQUOTATION is. Note that strictly speaking this principle is implausible. For example the quantifier 'everything' is sensitive to a contextually salient domain, so utterances of this sentence where a different domain is salient will result in something else being said. I think it is possible to control for this by introducing a context insensitive quantifier stipulated to range universally over all propositions. The possibility of doing this is not uncontroversial in this context (see Glanzberg [19] and Parsons [30]) but it is an assumption that many, including Williamson, share. Another issue is that presumably the word 'says' is context sensitive in the ordinary sense; I think this worry gets closer to the heart of the problem, and I pursue this in the later sections.

[33] See, for example, Goldstein's paper 'A Consistent Way with Paradox' [20] which makes several implicit appeals to this principle. It is worth contrasting the no-proposition theorists, who typically accept Moderate Disquotation, with the indefinite extensibilists mentioned earlier, who have the resources to deny it.

The idea that utterances of 'everything Alice says at t is untrue' say what you'd expect, if they say anything at all, is what makes options (1) and (2) more attractive than (3). But if Moderate Disquotation has failures the third option begins to look more attractive: if you're going to have to countenance mysterious propositions *anyway* you might as well go for the view that deviates least from orthodoxy.

3.2 Moderate disquotation

We have considered two kinds of no-proposition theory. The non-contextual version was subject to the problem that they have difficulties expressing their own view: if u is a paradoxical utterance of the sentence 'u is not true', then it's part of the non-contextual no-proposition view that u is not true (because it doesn't express a proposition). The natural way to express this is by uttering an instance of the same sentence, 'u is not true'. But on this view all utterances of this sentence fail to say anything. In this section we shall argue that the contextual version of the no-proposition view is also subject to this sort of problem.

Recall that an utterance is true if it says something true on the occasion on which it is made, and nothing it says is false. Here, as before, the apparent appeal to propositional truth and falsity in this definition is eliminable in favour of quantification into sentence position: in Prior's language, u is true iff $\exists P(Say(u,P) \wedge P) \wedge \forall P(Say(u,P) \rightarrow P)$ where $Say(u,P)$ means 'u was used to say that P'.

Consider now the sentence $L =$ 'no utterance of L both says something true and nothing false'.[34] Generalising moderate disquotation to this sentence gives us the following:

> If an utterance of 'no utterance of L says something true and nothing false' says anything at all it says that no utterance of L says something true and nothing false.

Thankfully this principle does not collapse into inconsistency: to get an inconsistency you'd have to assume that some utterance of L said truths and only truths. You can get around the paradox by maintaining that there are no utterances of L like this perhaps by insisting that no utterance of L says anything at all. In fact, if you properly formalize the moderate disquotational principle above in terms of propositional quantification, you can rigorously prove from this the result that there are no true utterances of L (utterances which say truths and only truths) in propositionally quantified classical logic:

Theorem From the above instance of moderate disquotation one can prove (in propositionally quantified logic) that no utterance of L says something true and nothing false.

I put the argument in a footnote.[35] Let me give you the gist of this argument informally:

According to these theorists the word 'true' is short for 'expresses some true proposition', and sentences involve the expression 'some proposition' can express different things in different contexts. Thus, unless we are in Alice's context when we utter Moderate Disquotation, we could end up expressing a falsehood. Note, finally, that Moderate Disquotation is vacuously accepted by non-contextual non-proposition views, for such views maintain that no utterance of 'everything Alice says at t is untrue' says anything.

[34] This is related to a fairly standard revenge problem for no-proposition token-based theories. Hazen [24], for example, considers the sentence L='Every sentence token equiform with L is false' (utterance falsehood has been substituted for something more explicit in my version). Williamson discusses the variant $L =$ 'L is not true in any context'. The following is an attempt to explicitly formalise these arguments and show that they in fact have exactly one premise (other than propositionally quantified classical logic): an instance of moderate disquotation.

[35] To formalize this properly we would need a binary expression $Say(u,P)$ which takes both a singular

1. **(Premise)** If an utterance of 'no utterance of L says something true and nothing false' says anything at all, it says that no utterance of L says something true and nothing false. (The premise is just an instance of Moderate Disquotation.)

2. If an utterance of L says something true and nothing false, it says something.

3. If an utterance of L says something, it says that no utterance of L says something true and nothing false. (from 1, the fact that L='no utterance of L says something true and nothing false' and Leibniz's law.)

4. If an utterance of L says nothing false and says that no utterance of L says something true and nothing false, then no utterance of L says something true and nothing false.

5. So, if an utterance of L says something true and nothing false, then no utterance of L says something true and nothing false. (from 2-4)

6. No utterance of L says something true and nothing false. (From 5.)

Assuming classical logic and the classical logic of the propositional quantifiers, the inference to (6) with (1) as the only premise can be formalized and proven as in footnote 35. Note in particular that all invocations of propositional truth and falsity ('says a truth', 'says nothing false', etc) can be eliminated in favour of quantification into sentence position.

So far so good: if we accept moderate disquotation we know that no utterance of L is true (i.e. says something true and nothing false.) (6) is an interesting theorem, and anybody who endorses the view under consideration should go about asserting (6). But there is a problem: theorem (6), just is the sentence L itself! Thus anyone who tries to utter (6) will have made an utterance which isn't true: an utterance of (6) just is an utterance of L, and as we just established, no utterance of L is true in the sense that it says truths and no falsehoods. The situation is just as dire as with the non-contextualist no-proposition view – when the contextualist no-proposition view attempts to describe her view she either fails to say anything with her utterances or she says falsehoods.[36]

---

term and a sentence to produce a sentence, and a binary relation $Utt(u, S)$ which tells us when an utterance u is an utterance of a sentence S. We use the variable u to indicate a quantifier is being restricted to an utterance of L (so e.g., $\exists u\phi$ is just short for $\exists x(Utt(x, L) \wedge \phi)$). Let Q be short for the conclusion $\neg\exists u(\exists P(Say(u, P) \wedge P) \wedge \forall P(Say(u, P) \rightarrow P))$. The instance of moderate disquotation we are interested may be formalized as $\forall u(\exists PSay(u, P) \rightarrow Say(u, Q))$. The argument then proceeds as follows

1. $\forall u((\exists P(Say(u, P) \wedge P) \wedge \forall P(Say(u, P) \rightarrow P)) \rightarrow \exists PSay(u, P))$. (Theorem of the logic of propositional quantification inside scope of $\forall u$.)
2. $\forall u(\exists PSay(u, P) \rightarrow Say(u, Q))$. (Premise).
3. $\forall u((\exists P(Say(u, P) \wedge P) \wedge \forall P(Say(u, P) \rightarrow P)) \rightarrow Say(u, Q))$. (From 1 and 2 by transitivity of the conditional.)
4. $\forall u(\forall P(Say(u, P) \rightarrow P) \rightarrow (Say(u, Q) \rightarrow Q))$ (universal instantation inside scope of $\forall u$.)
5. $\forall u((\exists P(Say(u, P) \wedge P) \wedge \forall P(Say(u, P) \rightarrow P)) \rightarrow Q)$. (From 3 and 4 by propositional logic inside scope of $\forall u$.)
6. $(\exists u\exists P(Say(u, P) \wedge P) \wedge \forall P(Say(u, P) \rightarrow P)) \rightarrow Q$. (From 5 by quantificational logic.)
7. $\neg Q \rightarrow Q$. (From 6 by definition of Q.)
8. Q. (7 by propositional logic.)

[36]It should be noted that strictly speaking quantifier phrases like 'no utterance' are context sensitive in the ordinary sense, so one has to be a little bit careful how one formulates these instances of Moderate Disquotation. One could, for example, artificially introduce an unrestricted quantifier into the language which,

The reader should note that unlike the unrestricted principle, Disquotation, the moderate version is not inconsistent with the assumption that certain utterances have been made, although endorsing it involves believing that no utterance of L is true. I have therefore, of course, overlooked a coherent position that involves believing that no utterance of L is true but refraining from ever uttering the sentence which appears to express this belief. I think it is an unattractive position to be in, not because it is inconsistent, but because it is hard to communicate it — people can certainly believe it, but they can't express their beliefs by uttering (6), or things that seem to entail (6) like 'no utterance of L says anything' or 'every utterance of (6) says at least one falsehood'.

Perhaps such a theorist might go ahead and utter (6) anyway, in a Wittgensteinian attempt to get across (or 'show') that no utterance of (6) says something true and nothing false, even if one does not succeed in saying anything. After all, even if in uttering (6) one fails to say anything, it seems clear what proposition you were intending to communicate, and were representing yourself as believing: that no utterance of L says something true and nothing false. But such responses forget the generality of Prior's theorem. S can be interpreted as 'Alice intended to communicate that', or 'Alice represents herself as believing that' so parallel arguments involving these operators show that the no-proposition theorist cannot even intend to communicate, or represent herself as believing only truths by uttering the variant of (6) that they are committed to in these parallel arguments.

So far we have just been discussing the no-proposition view. What of the multiple proposition view (i.e. option two): the view that Alice says two things, one of which is the proposition that everything Alice says at t is untrue? To make sense of this view we must be understanding 'says' in the loose rather than the strict sense.

Note, firstly, that this view doesn't avoid the mysterious proposition – there is still at least one other proposition that Alice says at t, and it's not the proposition that everything Alice said at t is untrue. More importantly, note that the view still seems to accord with Moderate Disquotation. In this case the restriction to utterances that say at least one thing becomes vacuous and we get to say something stronger:

> Anyone who utters the sentence 'everything Alice says at t is untrue' thereby says that everything Alice says at t is untrue (possibly along with other things.)

This might at first seem like a point in its favour, as moderate disquotation admittedly has a great deal of intuitive pull. Of course, it is also its undoing since the above paradoxes we have been discussing apply to any view that endorses moderate disquotation. In this case we can infer, as before, that no utterance of L is true in the sense that it is used to say only true propositions. In this case, although every utterance of L will say something, all utterances of L will say at least one falsehood (and possibly also a truth in some cases.)

in every context, allows one to quantify unrestrictedly over all utterances. Note that this extremely mild constraint certainly doesn't commit us to anything as controversial as a completely unrestricted quantifier or a universal domain. The class of utterances in our universe is plausibly finite. Unlike certain large classes, nobody would seriously deny that this class forms a set. A version of the indefinite extensibilist idea – that in any context one can always find a context in which more sets exist – doesn't really extend to concrete entities such as utterances: these things can't just pop into existence with a more expansive use of our quantifiers. It is a straightforwardly empirical matter which and how many utterances there are.

More worrying is the potential context sensitivity of the propositional quantifiers. Indefinite extensibilists insist that propositions are indefinitely extensible in the way sets are, so that it is not possible to introduce a context insensitive quantifier quantifying unrestrictedly over all propositions. Such a view has the means to resist our argument, although they fall afoul of other expressive problems (I shan't rehash these problems here, but see Williamson [44] section V).

This seems just as bad, since in order for the theorist to report this theorem she will have to utter the sentence ‘all utterances of L are untrue’, and will have thereby said a falsehood.

The multiple proposition theorist might utter (6) on the grounds that, although they ultimately end up saying falsehoods with utterances of (6), they also say the important truth that everything Alice says at t is untrue. We might think that this is the real truth that we intended to say by making this utterance, and that we represent ourselves as knowing when we make the utterance. The idea that we can intend to communicate only some of the propositions that are said with an utterance is subject to exactly the same problems we raised for the no-proposition view: it falls afoul of other instances of Prior’s theorem.

3.3 A Williamsonian rejoinder?

Although the conclusion of the last section poses a trouble for many contextualist views, Williamson makes a distinctive further claim about the semantic paradoxes that might in principle help make this conclusion sound better. According to Williamson ‘reflection on [the paradoxical] sentence causes normal speakers to give ‘say’, ‘true’ and ‘false’ a new set of correlative meanings, much like the previous ones except that the sentence in question says something in the new sense of ‘say”.

Here is why this might help. Suppose that Bob is about to reflect upon the sentence L, and conclude that no utterance of L is true by making an utterance of L itself. As we pointed out Bob will not succeed in saying anything. However, according to Williamson, once Bob has gone through this type of reasoning the words ‘says’, ‘true’ and ‘false’ shift their meanings.[37] For the sake of argument suppose they now denote the relation of saying$_+$ and the properties of truth$_+$ and falsity$_+$. So while Bob doesn’t succeed in saying anything with his utterance, perhaps he does succeed in saying$_+$ something, and although his utterance isn’t true, it may well be true$_+$. Note also that even though he doesn’t say anything and his utterance is not true, after the change has occurred we can truthfully report these facts using sentences like ‘Bob said something’ and ‘Bob’s utterance was true’, in the sense that these utterances will be true and true$^+$ at times after the change has occurred.[38]

So although the Williamsonian can’t express their view that no utterance of L says anything using the sentence ‘no utterance of L says anything’, they are in a position to express$^+$ their view using this sentence. Why should we care about this result? Well, provided expressing$^+$ and cognate notions play roughly the same role in the communication of beliefs among language speakers as expressing and cognate notions did, then the fact that we can express$^+$ the view seems like a perfectly good response to the objection that no-proposition theorists can’t communicate their view. This response therefore rests (although this is not something that Williamson emphasizes) on an instance of the thesis I called Semantic Pluralism earlier: that there is more than one natural language-world relation playing the ‘expressing’ role that takes a central place in semantic theorizing.

[37]This feature of Williamson’s approach distinguishes him from other no-proposition accounts.

[38]For the purposes of this discussion I have supressed some features of of Williamson’s account. For example, Williamson is using the word ‘says’ as a relation between sentences, contexts, languages and propositions; the expression ‘semantically expresses’ would be a more apt name for this notion. I am using the word in its ordinary sense, as it appears in speech reports like ‘so and so said that P’, which is not a language or context relative notion; the notion of utterance truth and utterance saying are similarly not language relative, as they are defined in terms of this notion. For Williamson reflection upon the semantic paradoxes brings with it a change of language. Note that this makes no difference as far as my language-independent notions of saying and truth are concerned: whatever language Alice is speaking, her utterance is true if for some P she said that P in virtue of having made that utterance, and P.

It will be instructive to consider a toy model of Williamson's account. I make no claim that the following represents Williamson's views accurately or completely. Indeed, several decisions need to be made in order to have a concrete view in sight, decisions which are not settled by Williamson's remarks. The following is thus just one of several ways of fleshing out the picture.

Assume there are three days: day 1, day 2 and day 3 during which exactly one shift in the meaning of the word 'says' occurs per day:

> On day 1, 'says' has a certain meaning, the relation of saying$_1$, on day 2 it means saying$_2$, and on day$_3$ it means saying$_3$.

Williamson talks about the subsentential meaning of the expressions 'says' and 'true' as though such a notion of meaning is unproblematic and unambiguous. But if the if there is more than one sentential meaning relation (saying$_1$, saying$_2$ and saying$_3$), we must take seriously the idea that there are multiple subsentential meaning relations, meaning$_1$, meaning$_2$ and meaning$_3$, each related by the principle of compositionality to their coindexed notion of saying. By drawing these distinctions, we muddy the relation that the word 'says' apparently has to the relation says$_1$ on day 1, says$_2$ on day 2 and says$_3$ on day 3. Could the word 'says', for instance, mean$_1$ the saying$_1$ relation on day one, whilst simultaneously meaning$_2$ some other relation? To avoid the dizzying effects of too much semantic pluralism, and to allow us to take Williamson's talk of *the* meaning of 'says' changing at face value, we will assume that subsentential meaning$_1$ is in fact the very same relation as subsentential meaning$_2$ and meaning$_3$ (allowing us to effectively omit the subscripts, if we wanted to).

The identity of subsentential meaning$_1$, meaning$_2$ and meaning$_3$ apparently conflicts with the distinctness of sentential saying$_1$, saying$_2$ and saying$_3$: a paradoxical utterance made on day 1 will not say$_1$ anything, whilst simultaneously saying$_2$ something. The differences between what is said$_1$ and said$_2$ cannot be due to differences in the subsentential meaning$_1$ and meaning$_2$ of 'says' (they are identical), and therefore must be due to a failure of compositionality: an utterance on day 1 might say$_2$ something without saying$_1$ anything, even when their subsentential parts have the same meanings$_1$ as meanings$_2$ on day 1.

In addition to being a form of semantic pluralism, Williamson's view is also broadly contextualist. The subsentential meaning of words like 'says' and 'true' changes between days. More explicitly, while the meaning$_1$, meaning$_2$ and meaning$_3$ of 'says' is the same on any given day, it changes between days, in unison, as indicated in the following table:

'says'	means$_1$	means$_2$	means$_3$
day 1	saying$_1$	saying$_1$	saying$_1$
day 2	saying$_2$	saying$_2$	saying$_2$
day 3	saying$_3$	saying$_3$	saying$_3$

Of course, the fact that the meanings$_i$ of 'says' change in unison is just guaranteed by our assumption that means$_1$=means$_2$=means$_3$ for subsentential meaning.

This fact allows us to make the welcome prediction that Williamson's diagnosis of the paradoxes is true$_i$ ($i = 1, 2, 3$) whichever day he makes the diagnosis. That is, utterances of "says' changes its meaning between day 1 and 2 and between day 2 and 3' are true$_i$ ($i = 1, 2, 3$) whether those utterances are made on days 1, 2 and 3. (Note that, without the assumption of the identity of subsentential meaning$_1$ with meaning$_2$ and meaning$_3$, the diagnosis offered by utterances of this sentence changes depending on what day it is made: for on day 1, you will being saying$_i$ (for any i) that 'says' changes its meaning$_1$ between day 1 and 2, and between day 2 and 3, but on day 2 uttering the same sentence would result in you saying$_i$ that 'says' changes its meaning$_2$ between day 1 and 2, and between day 2 and

3. If meaning$_1$ and meaning$_2$ are different, then these would be entirely different diagnoses. The ability to give the same diagnosis from day to day thus gives us independent reasons to identify the subsentential meaning relations.)

To see the contextual effects in action we need to look at different utterances of the *same* sentence. We can get a good picture by restricting our attention to nine utterances, u^i_j for $i, j = 1, 2, 3$, where u^i_j is an utterance made on day i of the sentence 'u^j_j is not true' (thus, for example u^1_1 is a self-referential utterance of the sentence 'u^1_1 is not true', made on day 1, whereas u^2_1 is a non-self referential utterance of the same sentence, made on day 2, about the day 1 utterance). Their behavior may be summarized in this table:

	says$_1$	says$_2$	says$_3$
u^1_1	Nothing	that u^1_1 is not true$_1$	that u^1_1 is not true$_1$
u^2_1	that u^1_1 is not true$_2$	that u^1_1 is not true$_2$	that u^1_1 is not true$_2$
u^3_1	that u^1_1 is not true$_3$	that u^1_1 is not true$_3$	that u^1_1 is not true$_3$
u^1_2	that u^2_2 is not true$_1$	that u^2_2 is not true$_1$	that u^2_2 is not true$_1$
u^2_2	that u^2_2 is not true$_2$	Nothing	that u^2_2 is not true$_2$
u^3_2	that u^2_2 is not true$_3$	that u^2_2 is not true$_3$	that u^2_2 is not true$_3$
u^1_3	that u^3_3 is not true$_1$	that u^3_3 is not true$_1$	that u^3_3 is not true$_1$
u^2_3	that u^3_3 is not true$_2$	that u^3_3 is not true$_2$	that u^3_3 is not true$_2$
u^3_3	that u^3_3 is not true$_3$	that u^3_3 is not true$_3$	Nothing

The first three rows concern three different utterances of the same sentence 'u^1_1 is not true'. u^1_1 can't say$_1$ anything: it was made on day 1, when 'true' means truth$_1$, and so if it said$_1$ anything it would say$_1$ that u^1_1 isn't true$_1$, from which it would follow (given the analogue of Utterance Truth for truth$_1$ and saying$_1$) that it was true$_1$ if and only if it wasn't. There is no similar barrier to it saying$_2$ or saying$_3$ anything: because on day 1 'true' means$_i$ ($i = 2, 3$) truth$_1$, compositionality tells us that it says$_2$ and says$_3$ that u^1_1 is not true$_1$, explaining the second and third entry of the first row. Because the meaning$_i$ of 'true' is truth$_2$ on day 2, and truth$_3$ and day 3 we can similarly explain rows two and three by compositionality. The next three rows concern 'u^2_2 is not true' and the final three 'u^3_3 is not true', and their explanations are completely analogous to the first three rows.

There is a final modeling choice, which we have not remarked upon, concerning whether the failure of an utterance to express a proposition only occurs when it has to (i.e. in cases that lead to paradox), or not. In the above we have entered 'Nothing' only in the cases where inconsistency would otherwise arise.

Williamson's discussion strongly suggests that the new meanings of 'says' and 'true' are more expansive than the old. He writes, for example, that 'encountering a semantic paradox might prompt us to enlarge what we mean by 'say'' (p15). As we can see, this is incompatible with the idea that we get failures of compositionality only when we have to. For example, it's completely consistent that u^2_2 says$_1$ something, and this is reflected in our table. But u^2_2 cannot say$_2$ anything on pain of paradox. Thus even though saying$_2$ is a later meaning for 'says', it is not more comprehensive than saying$_1$ (of course, by considering u^1_1 we see that saying$_1$ is not more comprehensive than saying$_2$ either). To capture the idea that the later meanings of saying are more comprehensive, we could artificially stipulated that u^2_2 and u^3_3 don't say$_1$ anything, and that u^3_3 doesn't say$_2$ anything. The entire table would thus instead look like this.

	says$_1$	says$_2$	says$_3$
u_1^1	Nothing	that u_1^1 is not true$_1$	that u_1^1 is not true$_1$
u_1^2	that u_1^1 is not true$_2$	that u_1^1 is not true$_2$	that u_1^1 is not true$_2$
u_1^3	that u_1^1 is not true$_3$	that u_1^1 is not true$_3$	that u_1^1 is not true$_3$
u_2^1	that u_2^2 is not true$_1$	that u_2^2 is not true$_1$	that u_2^2 is not true$_1$
u_2^2	Nothing	Nothing	that u_2^2 is not true$_2$
u_2^3	that u_2^2 is not true$_3$	that u_2^2 is not true$_3$	that u_2^2 is not true$_3$
u_3^1	that u_3^3 is not true$_1$	that u_3^3 is not true$_1$	that u_3^3 is not true$_1$
u_3^2	that u_3^3 is not true$_2$	that u_3^3 is not true$_2$	that u_3^3 is not true$_2$
u_3^3	Nothing	Nothing	Nothing

In this second model, compositionality fails even when we can consistently assume it to hold.

Let us discuss some of the features of the view I've just sketched. Firstly, the temporal element of the view isn't as important as it might seem at first. We've seen that it isn't really essential to the view that the meaning of 'says' becomes more expansive over time, and we've sketched a non-expanding version of the view in which saying$_1$, saying$_2$ and saying$_3$ are all pretty much isomorphic, but for differing over which utterance, u_1^1, u_2^2 or u_3^3, they fail to relate to a proposition.[39]

Secondly, and more importantly, the sense in which saying$_2$ comes in to play later than saying$_1$ is quite thin. On day 1, when we make utterances of ordinary sentences we simultaneously say$_1$, say$_2$ and say$_3$ something (there will be some gaps regarding what we say$_1$/say$_2$/say$_3$ with paradoxical utterances, but this is the general picture). Moreover, we suggested, in response to the expressive worry raised in the last section, that the three different saying relations are all equally important for describing how we communicate; it's not OK to say$_2$ false things on day 1 any more than it is to say$_1$ false things. The view therefore seems to be one in which utterances on any given day have three important semantic properties; the three semantic statuses of these utterances do not vary from day to day. (The thin sense in which saying$_2$ comes in to play later than saying$_1$ is simply that it is the meaning of 'says' on day 2, and not day 1.)

But the view has some unintuitive elements as well. For example, it requires a quite remarkable and mysterious connection between the meaning of certain words and the practices of a certain group of people consisting mostly of logicians and philosophers (the group of people responsible for doing the most reflecting upon paradoxical sentences.) Puzzling questions quickly present themselves: if two people reflect upon a paradoxical sentence in quick succession, the language will successively undergo two quick changes of meaning. However if they reflect upon the sentence at exactly the same time, will the language undergo two changes or one? One might also wonder whether merely thinking about the paradoxes induces the change, or must it be aloud? If it is merely thinking then would it be possible for a non-English speaker to change the meaning of English words just by thinking about the paradoxes? What if, for example, an alien was aware that Alice had attempted to say that everything she's saying at t is false, whilst being completely unaware of how she tried to say it and of the English language quite generally. Could such a person really have an effect on the meaning of English words?

[39] It's also worth noting that in our non-expanding model, we can make many of the predictions Williamson achieves by having only two saying relations, with the meaning of 'says' alternating between when we reflect on paradoxical utterances. For by inspecting the entries of our table we see that on day 3, saying$_1$ and saying$_3$ are in agreement about the immediately preceding utterances (i.e. the day 2 utterances), and every proposition in our table may be said$_1$ or said$_2$ with an utterance of one of the three sentences in question.

Secondly, the view I have described isn't compositional — this is at least a *prima facie* cost. Indeed, if the explanation of our failing to say anything with an utterance boils down to a failure of compositionality, a sort of revenge paradox arises quite straightforwardly. Let the *compositional closure* of some subsentential meaning relations, written 'says^{cc}', be defined as follows:

> An utterance of S says^{cc} that P iff the meanings of the parts of S, when applied to each other in accordance with the syntactic make up of S, compose to the proposition that P.

Even if an utterance fails to say something, provided its subsentential parts having meanings, it will still bear the compositional closure of saying, saying^{cc}, to some proposition. Instances of Prior's theorem can be stated that instead use the saying^{cc} relation.

Indeed, these sorts of reflections have lead me to think that the best way to avoid the sentential disquotational principles that lead to paradoxes is not to deny compositionality at all. A broadly contextualist view, like Williamson's, already denies disquotational principles as they apply at the subsentential level (such as 'says' means says, 'true' means true, and so on). Once one has made this concession, however, one has no need to give up compositionality to avoid the problematic disquotational principles. We shall develop this view in the next section. But first, a small digression.

3.4 What it sounded as though Alice said

As we mentioned in section 2.2, there are other puzzling instances of Prior's theorem which the no-proposition theory doesn't have much to say about.

Consider again reading S in Prior's theorem as 'at t it sounded as though Alice said that', and suppose moreover, that at t Alice utters the sentence:

> Anything it sounded as though I said at t is false.

Anticipating a paradox parallel to the one we have been discussing, the no-proposition theorist might insist that Alice simply didn't succeed in saying anything with her utterance. However, this response does not avoid the relevant paradox, for it entails that if it merely *sounded* as though Alice said that everything it sounded as though she said at t is false, then there are two things it sounded as though she said at t. The parallel response to this instance of Prior's theorem is to insist that there is nothing it even *sounded* as though Alice said t. But, whether or not Alice said anything at t, it seems self-evident that it at least *sounded* as though she said something.

Now this is not a direct objection to the no-proposition theory, since it did not set out to solve every paradox that could be considered an instance of Prior's theorem. But it does highlight a deeply puzzling phenomenon for which the no-proposition theory offers no solution. One might expect a fully satisfying response to Prior's paradox to also explain this variant.

4 Radical Anti-Disquotationalism

The above considerations cast an unfavourable light on the no-proposition and multiple-proposition interpretations of Prior's theorem – that Alice said nothing in uttering her sentence or that she said several things: both positions have trouble expressing their views.

This leaves only the third option: That Alice says exactly one thing when she utters the sentence 'everything Alice is saying at t is untrue', but it is not the proposition that everything Alice is saying at t is untrue. This response, however, immediately raises the question: *what on earth* is the proposition that she said in that circumstance, if not the proposition that everything she said at t is untrue?

In order to explain this puzzle, and much more, I will defend and appeal to the following thesis:

Semantic Pluralism There is a large number of language-world (and mind-world) relations that are all extremely similar to one another, each one playing similar roles in the theory of communication, and none more deserving of the scrutiny of semantic theorizing than any other.[40]

By a language-world (mind-world relation) I simply mean a relation holding between utterances (or mental tokens) and propositions. The kind of relation you might report as standing between an utterance and a proposition P when the producer of the utterance has said that P, or a belief (or any other attitude) and a proposition P if it is a belief that P. I shall focus on the linguistic case in what follows. In light of the discussion in section 3.3 it is natural to view Williamson's view as a version of semantic pluralism.

I claim that although semantic pluralism is quite a natural view it entails the following much more surprising thesis:

Radical Anti-Disquotationalism Hardly any of these language-world relations relate utterances of 'snow is white' to the proposition that snow is white.

In fact, I take Radical Anti-Disquotationalism (RAD, for short) to entail a long conjunction of surprising theses like the one above. In general one must resist the urge to narrow down the relations that play the expressing role to those which 'behave disquotationally' (I put this expression in scare quotes because it will soon become clear that 'behaving disquotationally' is not even a well defined property).

I choose the sentence 'snow is white' because it is often taken to be a paradigm case of a sentence that satisfies the disquotational schema. Actually the sentence 'snow is white', like most English sentences, is vague and therefore, at least according to one orthodoxy, context sensitive.[41] Since context sensitive sentences do not straightforwardly satisfy disquotational principles, and vagueness is so pervasive, we might already find ourselves to be committed to pervasive failures of disquotation. However context sensitive sentences still satisfy qualified versions of disquotational principles — for example, one might insist that utterances of 'snow is white' made in the present context say that snow is white. My arguments for RAD will establish that even these qualified disquotational principles have failures.

Why be interested in radical anti-disquotationalism? While it may be interesting and surprising in its own right, it also encodes exactly the kind of assumption that was responsible for the paradoxes we have been discussing. Once we've accepted a natural picture in which disquotational principles are in general false and fail in paradigm cases of ordinary semantic theorizing, we can see that the semantic paradoxes are actually quite unexceptional and require no revision to our semantic practices. This is shown more rigorously in section B where a model of semantic pluralism is given of a language which can reason about its own semantics in a completely general way.

[40] Semantic pluralism is not the same as alethic pluralism, the view that different subject matters require different truth predicates.

[41] Although, for a contrasting opinion, see Cappelen and Lepore [10].

It should be stressed that although RAD encodes an important sense in which the present view is 'anti-disquotational', there are other kinds of disquotational ideas one might want to give voice to. In particular, RAD does not entail that it's always bad to make utterances of disquotational *sentences*. Indeed it's consistent that utterances of the sentence ' 'snow is white' means that snow is white' are related by these language world relations only to truths. We discuss this point further in section 5.2.

4.1 Semantic Pluralism

The business of semantics is to match up bits of language to bits of the world; names to objects, predicates to properties, sentences to propositions and so on. If this was all that was required, matchings of this sort would not be hard to come by – there are lots of uninteresting ways of matching up utterances to propositions, for example. To fall within the purview of semantics a relation between an utterance and a proposition must additionally play some role explaining how that utterance causes beliefs in that proposition amongst speakers of the relevant language. Semantic pluralism is simply the view that there are many relations of this sort that fit the bill.

It would, I think, be incredibly surprising if there was exactly one relation between utterances and propositions that satisfied that kind of communicative role (described in more detail below). Thus semantic pluralism should have a considerable amount of prima facie appeal. However, it is worth examining the arguments for semantic pluralism in more detail to get a view of the mechanisms by which it comes about.

A central theoretical term for many philosophers and semanticists concerned with meaning is that of a sentence *semantically expressing* a proposition in a language relative to a context. This is a technical notion, although its can be pinned down by its wider theoretical role. It's unlikely that this role can be completely specified by a finite list of easy-to-state principles, however the following remarks capture some of its distinctive features:

1. We clearly use sentences to communicate with one another, yet what a sentence semantically expresses in a context will in general be much weaker than the total information imparted by a given use of that sentence in that context. When I utter a sentence my audience learns that I have vocal chords, a British accent, that I believe what I said, as well as any conversational or conventional implicatures my utterance might carry. None of these are part of what the sentence semantically expresses in that context.

2. What a sentence semantically expresses in a context often coincides with *what the speaker said* when a speaker makes an utterance of that sentence in that context.

3. If a speaker said that P with an utterance of the sentence S, and it is moreover the case that P, then that utterance is true; in these cases one can talk of the *sentence* being true in that context in the language in question. Thus given 2, we can say that a sentence S is true in L at c if and only if for some P, S semantically expresses P in L at c, and P.[42]

4. One shouldn't assert false propositions. Since one way to do that is to assertively utter a sentence which expresses a false proposition, it follows that assertive utterance is nor-

[42] Earlier in the paper I defined utterance truth as 'saying something true and nothing false'. However if we understand saying and expressing in the strict sense it follows that things say and express no more than one thing. Thus we can drop the clause 'and nothing false' from our definition.

matively constrained by the semantic notions of expression and truth: one shouldn't assertively utter a sentence if it is not true, or if it expresses a false proposition.

5. The relation of semantic expressing is governed by a principle of compositionality; for example a disjunction expresses the disunction of the propositions each disjunct expresses. One can generalise this to the subsentential principles with a generalised notion of semantic value.

These remarks, I take it, encapsulate some of the most important aspects of the role that technical notions like 'semantic value', 'semantic expression', 'truth' and so on, play in philosophical discussions of semantics and in linguistics.[43]

Earlier we made the simplifying assumption that technical talk of a sentence 'semantically expressing' a proposition P in a context can be eliminated in favour of just talking about what an agent uttering that sentence in that context would have said. Although this leaves talk of 'semantic expression' as loose and context sensitive as the notion of 'what is said', this is not entirely inappropriate: it would be remarkable if philosophers had managed to precisely and context insensitively latch onto a singly important relation in the course of their theorizing with phrases like 'semantically expresses'.[44] However the crucial point, which everyone can accept, is the intimate connection the technical notions have have with the pretheoretic notion of *what is said* by an utterance of a sentence.

Let me begin with a suggestive, although not decisive motivation for semantic pluralism. As noted already, speech reports of the form 'so-and-so said that P' are notoriously context sensitive. If Bob, for example, says that he will be up at 5am tomorrow then in some contexts I can report him by saying 'Bob said that he would be up early' (say, if we want to know if he'll be up before noon), whereas in others I cannot, and might even be able to say 'Bob did not say that he would be up early' (for example, if we were in a court of law determining whether Bob had committed purjury.) Thus there is a plethora of distinct but closely related relations that the word 'says' can be used to express, and for each of these conceptions of what it is that a person said with a given utterance there is a corresponding semantic relation relating that utterance to that proposition. Whence, semantic pluralism.[45]

Now the above example employed a use of the word 'says' that was fairly loose – it related Bob to the proposition that Bob would be up early when usually we would describe him as having said something much stronger. Presumably the notion of 'what is said' that is employed in elucidating the notion of semantic expression is a bit more strict than that. However once we have conceded that the word 'says' can be used to mean lots of different things, many of which differ from one another only by tiny factors, it seems unlikely that even in philosophical contexts we are able to settle on a perfectly consistent and unique use of the word 'says'. More importantly, even if they could, by some tremendous feat, do this, it seems unlikely that they would be picking up an a uniquely important attitude. Many of the relations that the word 'says' expresses can feature in explanations of how people communicate – one can say things like 'A did this and that because B said so and so' in

[43] See Soames [38] chapter 4 for a more detailed description of this role.

[44] The eliminativist strategy is not quite so straightforward for the generalized notion of semantic value which applies to semantic categories other than sentences. I suspect we can make good sense of what a person is referring to with a use of a singular term (not to be understood as what they are intending to refer to.) However the slightly more technical notions seem to be needed if we are to extend talk of semantic value to predicates and other subsentential categories.

[45] If we adopt the eliminativist strategy suggested above, this conclusion follows directly from the context sensitivity of 'says'.

a number of different contexts, and explain A actions in terms of a number of different attitudes B holds towards a certain proposition.

While I take the argument from context sensitivity to be suggestive, it is far from watertight. In most contexts the word ‘says’ appears to relate people to more than one proposition at a time. It is thus consistent with everything I’ve said to have a picture in which each utterance is associated with a strongest proposition corresponding to the conjunction of the propositions said, along with a few of its obvious logical consequences.[46] One could insist that all context sensitivity in the word ‘says’ is context sensitivity deriving from how many consequences of that proposition we also count as having also been said. This view is consistent with the idea that there is a single proposition that is what an utterance has *strictly speaking* been used to say, and which all uses of the word ‘says’ relate that utterance to, and that any context sensitivity concerns only which of the logical consequences of that proposition come along for the ride. On this picture one could simply identify the proposition semantically expressed with the strongest base proposition that is constant between all uses of the word ‘says’; such a view does not obviously motivate semantic pluralism.

A better model of semantic pluralism is suggested by consideration of the vagueness of words like ‘says’ and ‘semantically expresses’.[47] As many people have observed, vague words like ‘tall’, ‘red’, ‘bald’ and so on, typically also exhibit a degree of context sensitivity which is distinctive in that the differences in meaning are only very slight. Moreover, the cloud of propositions that’s associated with a sentence like ‘Harry is bald’ is typically such that there isn’t a member of that cloud which entails all the other members of that cloud: there is no proposition in the cloud that could be a plausible candidate for what has strictly speaking been said.

Note, however, that even if vague words are not always context sensitive, it is surely true that when a term is vague there is also a plethora of of closely related denotations of the appropriate semantic type associated with that word, each of which playing roughly the same role in thought and communication as any other. For example even if we assume that the predicate ‘tall’ picks out exactly one property, there are plenty of closely related properties, differing only with regards to the minimum heights they are compatible with, each of which seem equally good candidates to be associated with the word ‘tall’. The word ‘says’ is surely the same in this regard. Although we do not have as clear a dimension as height to make the point, surely whether a person has said that P, or whether an utterance semantically expresses P, supervenes on the microphysical facts in some way, thus there will be a host of candidate relations that differ only slightly with regard to which microphysical states they are compatible with.[48]

Before I give some more concrete examples of such relations, let us pause to stress how general these points are. In the above arguments I have primarily focused on the vagueness and context sensitivity of the word ‘says’. Perhaps you think the technical notion of semantic expression should not be as closely tied to this pretheoretic notion of ‘what is said’ as I have assumed. Other conceptions of the central language-world relation are available; perhaps one can spell it out in terms of conventions of truth and trust amongst the linguistic community

[46]There is of course still the strict sense of ‘says’ that you can define as the conjunction of things said in the loose sense. However the data supporting the context sensitivity of ‘says’ all concern the ordinary use of ‘says’.

[47]The following discussion is heavily inspired by the discussion in Dorr and Hawthorne [13].

[48]See Dorr and Hawthorne [13] for a way of make the notion of a ‘slight difference in microphysical facts’ precise. Dorr and Hawthorne make a very similar argument to the the one above, arguing that “In almost all cases, an expressions actual meaning is surrounded by a vast cloud of slight variants which seem just as well qualified to be possible meanings.”

(Lewis [28]), or in terms of interpretations that maximise the number of true beliefs of the linguistic community (Davidson [11]) or in some other way. Although these accounts place more emphasis on the attitudes of belief and desire and less on the attitude of saying, the considerations we have been considering extend fairly naturally to these other attitudes. However we spell it out, it would be utterly wild to think that the resulting account will not involve words that are vague or context sensitive, and thus it seems incredibly natural to think that there will be more than one relation which plays the semantic expressing role according to these theories. Thus analogous conclusions can be drawn even under the assumption of these alternative pictures.

Let us now make the above ideas more explicit. Here is one way. It is a platitude that as languages evolve, words change their meanings – this much is true whatever we take meaning to be. Let us suppose that an expression E now has a different meaning than it did fifty years ago. Now whether this change happened gradually, with the expression continually changing its meaning by small amounts, or in a small number of big jumps, there must have been a moment of time at which the expression had a meaning which it didn't have one nano-second earlier. This change is naturally going to be a function of how the use of the expression is changing over time. Now presumably there is a very similar function of the use of E, calling this 'meaning*', according to which the change in meaning* is delayed by a nano-second relative to the change in meaning. Now, I maintain that it would quite remarkable if the thing I've presumptuously called 'meaning', and not the thing I've called 'meaning*', played a more important role in the theory of communication. For even if we focused exclusively on communication occurring at the nano-second at which they differ, it is hard to see how any important feature of communication could feature one of these relations more centrally than the other.[49]

Here is another reason to think there is a cluster of closely related language-world relations, this time stemming from the observation that there are typically a cluster of closely related propositions associated with each *non*-semantic sentence. For example, in addition to the proposition an utterance of 'snow is white' actually expresses, call that P, there are a cluster of closely related propositions that differ from one another only slightly. Perhaps the strongest thing P entails about the reflectiveness of snow is that snow reflects a particular range of the visible spectrum R, whereas the strongest thing entailed by another proposition in the cluster, Q, is that snow reflects a very slightly shifted range R'. Since, by assumption, neither the range R nor R' is contained within the other, neither P nor Q is entailed by the other (the cluster of propositions associated with 'snow is white' is not linearly ordered by entailment.) Now in addition to the relation that I've (again, presumptuously) called 'semantic expressing' that relates this utterance of 'snow is white' to P, there is another very similar relation that relates utterances of sentences involving the word 'white' to ever so slightly different propositions, and in particular relations that match this utterance to Q. Once again it would be remarkable if one of these two relations was theoretically more significant than the other, for the way in which beliefs that P pattern with the use of the sentence 'snow is white' is roughly the same as the way in which beliefs that Q pattern with this usage.

To fix ideas, it might be instructive to consider a toy model of closeness that allows us to say, a bit more precisely, what it means for (i) two propositions to be similar, and (ii) what it means for two expressing relations, represented by functions from sentences

[49]Note that even if we model the meaning of a sentence as a cluster of propositions rather than a single one, it's just as clear that a sentence can change the cluster of propositions it means over a period of time and an analogous argument can be run on that assumption.

to propositions, to be similar. Subsequent discussion will not depend on the mechanics of this specific model, but it is informative due to its concreteness. We will adopt a standard possible worlds framework in which propositions are identified with sets of worlds, properties with functions from individuals to sets of worlds, and so on (see appendix A). First, assume that we have some way of measuring similarity between two worlds. Formally, this may be represented by a metric function, d, taking two worlds and yielding a non-negative real number representing the extent of disimilarity between them, $d(x, y)$.[50] We will introduce the concept of *perturbation*. A perturbation determines two things: a permutation of worlds that maps worlds to similar worlds, which we represent formally by the following condition

$d(w, iw) \leq \alpha$ for every world w.

Here two worlds count as similar if their disimilarity does not exceed the threshold, α. Secondly, it determines a permutation of individuals mapping individuals to similar individuals, where individual similarity may also be capture by a metric (for example, there appear to be lots of distinct but overlapping hunks of rock in the vicinity of Mt. Kilimanjaro, differing over their exact boundaries).[51]

The notion of a perturbation gives us a natural way to say when two propositions are close: namely, if one can be mapped to the other by a perturbation — for some perturbation i, $iP = Q$, or $iQ = P$, writing iP for $\{iw \mid w \in P\}$. Likewise, two individuals are close if one can be to mapped to the other by a perturbation. More importantly, we can say what it means for a property F, as represented by a function from individuals to propositions, to be close to another. For every perturbation, i, has an inverse, i^{-1}, whose action on propositions and individuals is given by the inverse of is action on propositions and individuals. We shall say that perturbation i maps F to the function that maps an individual a to $i(F(i^{-1}(a)))$.[52] Thus we may say that one meaning relation, as represented by a function from individuals to propositions (mapping non-sentences to some fixed proposition, $\perp$ say), is close to another if there is a perturbation mapping one to the other in this sense. (Note that the action of a perturbation may be extended in this manner to any type whatsoever, since once we have said what a perturbation does to things of type σ and of type τ we may use the same idea to define the action of the perturbation on type $\sigma \to \tau$; this is made a bit more explicit in the appendix.)

The above model predicts the sorts of connections we'd expect given the forgoing considerations. For example, if m_1 and m_2 are similar meaning functions, then we should expect m_1 and m_2 to map the sentence 'snow is white' to similar propositions. For suppose that $i(m_1) = m_2$ for some perturbation, i. It's natural to think that certain mathematical and abstract objects like sentences, but unlike mountains, are fixed by perturbations: a sentence is sufficiently disimilar from any other object that no perturbation could map it to anything but itself. In which case it follows that $iS = i^{-1}S = S$, so that

[50]We assume that d is a *metric*: i.e. for any worlds x, y and z, $d(x, x) = 0$, $d(x, y) = d(y, x)$ and $d(x, z) \leq d(x, y) + d(y, z)$. Following [13], we might look to the underlying physics to find natural metrics: for example, in classical mechanics a world is uniquely determined by the position and momenta of each particle. The distance between two worlds (containing the same particles) may be identified by the Euclidean distance between two such vectors of quantities, in a given numerical representation. I will not assume any specific interpretation of d in what follows, however.

[51]Not all pairs of permutations will correspond to a perturbation. Intuitively, they must be coordinated: if, for example, a perturbation, i, maps a hunk of rock x to another, y, then i must map the proposition that x is exactly 19,341 feet to the proposition that y is exactly 19,341 feet.

[52]That this choice is natural needs some independent defense that would take us too far afield here, but see e.g. Bacon [5] for some of the relevant theory.

$m_2(S) = i(m_1)(S) = i(m_1(i^{-1}S) = i(m_2S)$, demonstrating that $m_1(S)$ and $m_2(S)$ are related by a perturbation, and are thus similar.

Let me end, then, by making two further comments about the generality of this discussion. Firstly, it is hard to resist the arguments above by maintaining that the semantic expressing relation relates each utterance not to a single proposition, but to the cluster of propositions itself. For the clusters themselves can come in clusters of similar clusters of propositions, differing only slightly over what they jointly rule out. In what follows I shall continue to assume that each of the semantic relations we are concerned with is *strict*: they relate each sentence to a unique proposition.

Secondly – although it is beyond the scope of this paper to explore this idea thoroughly – very similar arguments can be run for to mind-world relations by noting that attitude reports are also subject to similar considerations as speech reports.[53] For example, let b the be mentalese equivalent of the English sentence 'snow is white'. Suppose, by analogy with the linguistic case, that b is a belief that P, where P is a proposition that is inconsistent with certain hypotheses about the reflectivity of snow, and Q is a very similar proposition that is inconsistent with a slightly different range of hypothesis. A parallel with our earlier discussion arises, for there is another belief-like attitude belief*, in which b counts as a belief* that Q, which plays roughly the same role in explaining our actions and verbal behaviour as the attitude of belief did.

Here is some useful terminology for what follows. I shall use 'says$_1$', 'says$_2$', 'says$_3$' as place-holders for a complete spelling out of the various different utterance-proposition relations. The subscripts are not to be thought of as an argument place for a number, they are merely typographical ways of distinguishing the different relations. Corresponding to each of these relations are subsentential semantic relations: for singular terms I shall use 'refers$_1$', 'refers$_2$' and so on, and more generally 'means$_1$', 'means$_2$', ... for other semantic categories.

We shall suppose that each family of denotation relations for a given index satisfies the role spelled out in (1)-(5). Thus, for example, co-indexed denotation functions must satisfy the principle of compositionality with respect to one another: e.g. what an utterance says$_1$ is a function of what its parts denote$_1$ on that occasion of use (and similarly for what that utterances says$_2$, says$_3$ and so on). Some portion of this role concerns the penumbral connections between co-indexed notions. And since our arguments suggest that pluralism is quite pervasive for attitudes, one might suspect that many of the concepts involved in this role are pluralized, including belief, assertion, desire and so forth. One might worry that the expressing relation is as a result radically underdetermined by the expressing role — that our theory has been Ramsified to the point that it is tells us nothing. But note that other parts of the role do impose objective constraints, relating pluralized notions to unindexed concepts. For example, consider the maxim that one shouldn't assert that P when it's not the case that P. Presumably assertion, like saying, is pluralized so we actually get a family of constraints saying that we shouldn't assert$_i$ that P when not P, for each i. This relates the indexed notion of assertion$_i$ to the objective unindexed notion of propositional truth and falsity. Given (2) this also implies that it's OK to make an assertive utterance only if that utterance is true$_i$ for every i. For if you made an utterance that was false$_i$ for some i, you would thereby assert$_i$ something that was not the case, contradicting the relevant instance.

[53] Many people have noted the vagueness and context sensitivity of belief reports. Although some of these considerations stem from fairly specific considerations of Frege puzzles, others are quite general (see, for example, Field on mental representation [15].)

4.2 Semantic Pluralism entails Radical Anti-Disquotationalism

Although the paradoxes give us reason to reject particular instances of disquotational reasoning, here I argue that semantic pluralism entails far more radical violations of disquotational principles.

Radical Anti-Disquotationalism Hardly any of the cluster of important language-world relations relate utterances of 'snow is white' to the proposition that snow is white.

Of course, if 'snow is white' is context sensitive the idea that an utterance of 'snow is white' in context c says that snow is white is only plausible if c is a context relevantly like the context we are in right now. However my case for radical anti-disquotationalism does not rest on the context sensitivity of the sentence 'snow is white', for it will show that for any utterance of 'snow is white' (even ones made in this very context), most of the semantic relations will not relate that utterance to the proposition that snow is white.

The interest of radical anti-disquotationalism for our purposes is that it demonstrates that there is nothing exceptional about 'paradoxical' utterances and moreover it deflates a number of potential revenge arguments resting on weakenings of the disquotational principles.

It is instructive to look at a couple of instances of this argument: one for saying, and one for reference.[54] Suppose that says$_1$ and says$_2$ are two such semantic relations. Suppose, moreover, that Alice makes an utterance of 'snow is white' and that the two semantic relations deliver differing verdicts for Alice's utterance: says$_1$ relates Alice's utterance to the proposition that P and says$_2$ relations relates her to a distinct proposition, Q (where, lets say, P and Q differ marginally over the minimum fluffiness that something needs to be to be snow). Since at most one of P and Q can be identical to the proposition that snow is white, at most one of the two semantic relations will relate Alice to the proposition that snow is white. The thought generalises, for if disquotational principles fail roughly half the time when there are only two relations, when there are several relations which do not agree with one another about Alice's utterance at most one of them can relate Alice to the proposition that snow is white, and thus most of them will not relate Alice to the proposition that snow is white.

Parallel arguments establish similar conclusions for reference. There are plenty of mountain-like objects in the vicinity of Mt. Kilimanjaro that differ from one another only slightly over their boundaries — for example, some might include a small piece of dust near the bottom and others might not. Mt. Kilimanjaro is, of course, one of these objects, but there is nothing particularly special about it — it does not stand out as haloed among the other similar objects. There is presumably vagueness concerning the referent of 'Mt. Kilimanjaro'. It is natural to think there is also semantic multiplicity here — lots of extremely similar reference-like relations each relating the name 'Kilimanjaro' to one of these different mountain-like objects. I take it that none of these mountain-like objects has the unique privilege of being the most suitable semantic value of the name 'Kilimanjaro', not even Mt. Kilimanjaro, and that none of the variant reference relations plays a more important semantic role than any other. Yet since only one of these mountain-like objects is Mt. Kilimanjaro most of these reference relations will not relate the name 'Kilimanjaro' to Kilimanjaro.

It is initially quite tempting to turn this argument on its head and take it as a refutation of semantic pluralism. 'Look', one might argue, 'perhaps refers$_1$ and refers$_2$ play very similar roles, however if one of them is relating 'Mt. Kilimanjaro' to Mt. Kilimanjaro and the other

[54] Here I draw on the discussion in Hawthorne [23].

is not then surely the relation that does is the more important language-world relation, and we should not be concerned with the other relation in our semantic theorizing'. We may also apply this argument to rule out any saying relation that does not relate 'snow is white' to the proposition that snow is white; indeed, if we repeat this style of reasoning for every expression in the language we may show that if there are multiple important semantic relations, they must all agree with one another about which expressions are related to which denotations, showing them all to be extensionally equivalent. This, I take it, is enough to rob semantic pluralism of its interest. Let's call this the *disquotational argument against semantic pluralism*.

Quite apart from the semantic paradoxes, disquotational constraints like the one suggested above should be treated with caution. Consider the following parody argument:

> The word 'me' can be can be used in different contexts to pick out different people. In some contexts it picks out Alice, in some Bob and in others me. However there is clearly something theoretically distinctive about the last kind of context which makes it more deserving of the attention of semantic theorizing, for it is the only context in which 'me' behaves disquotationally.

The absurdity of this line of reasoning is dramatized by observing that Alice or Bob could have verbally made the exact same argument and arrived at very different conclusions.

Although semantic multiplicity is not the same as context sensitivity, the similarities are evident. The problem in the above argument was that I used (as opposed to mentioned) a word, 'me', that had several equally good referents associated with it; in the present case we used an expression, 'Mt. Kilimanjaro', that while not context sensitive is associated with many different individuals that are all on a par as potential referents for 'Mt. Kilimanjaro' go.

Of course, there is an apparent tension that needs to be defused. On the one hand, we have good reasons to think that no particular object located in that particular region of Tanzania is more special than any other, metaphysically, semantically, or otherwise. On the other hand, we know from first-order logic that exactly one object is identical to Mt. Kilimanjaro. Isn't being identical Mt. Kilimanjaro itself a source of specialness, a property which none of the other hunks of rock share? Or perhaps we should follow common sense and maintain that there is only one mountain in that part of Tanzania, and conclude that, actually, the different hunks of rock in that vicinity are not on a par — the one that's a mountain is special. Semantic Pluralism helps us see our way out of this tension. When we make speeches involving names like 'Mt. Kilimanjaro' we simultaneously say$_1$ and say$_2$ different propositions. Suppose that 'Mt. Kilimanjaro' refers$_1$ to a certain hunk of rock, Hunk$_1$, and refers$_2$ to another, Hunk$_2$, and that the principle of compositionality holds. Then when we make speeches like 'Mt. Kilimanjaro is the best candidate referent for 'Mt. Kilimanjaro'' (whatever 'best' might be taken to mean), we say$_1$ something about Hunk$_1$, namely that it is the best candidate referent for 'Mt. Kilimanjaro', or something close enough. And we simultaneously say$_2$, with the same utterance, something about Hunk$_2$, namely that it is the best candidate referent for 'Mt. Kilimanjaro'. Such speeches, thus, clearly do not favour Hunk$_1$ or Hunk$_2$. Moreover, they appear to be incompatible propositions, so the speech in question cannot be acceptable. Similar speeches appealing to the idea that there is only one mountain in that vicinity similarly misfire, as 'mountain' is presumably also related by different semantic relations to different but related properties, each containing a unique hunk of rock in its extension without agreeing about which hunk of rock that is.

The key to making ones peace with these anti-disquotational conclusions is to recognise that there is nothing special about Mt. Kilimanjaro over the other mountain-like entities in its vicinity, and that the proposition that snow is white is no more important than the plethora of propositions distinct from it but with very similar modal profiles. Once we have made our peace with that, we may also deflate the disquotational argument against semantic pluralism argument we made above. For suppose I make the following announcement:

> Relations that relate utterances of 'snow is white' to the proposition that snow is white are better than those that don't.

Since 'snow is white' says$_1$ that P and says$_2$ that Q, then we can ask what I've said$_1$ and said$_2$ when I make the above announcement. I've said$_1$ that relations that relate utterances of 'snow is white' to P are better than those that don't (or something like that) and I've said$_2$ that relations that relate utterances of 'snow is white' to Q are better than those that don't.[55] Not only is there a perfect symmetry here between P and Q, but I've said$_1$ and said$_2$ contradictory things, which explains why it might be bad to make announcements like this.

The core issue at stake here is whether there is something special about Mt. Kilimanjaro and the proposition that snow is white, over the other closely related objects and propositions in the vicinity. This is an extremely seductive thought, abetted by the naturalness of certain disquotational speeches, but a thought that I think must ultimately be resisted.

We must be vigilant in applying all of these morals to our subsequent theorizing involving semantic words. We have been theorizing in terms of the relation of *saying*, which we have argued is but one of a large family of similar language-world relations that satisfy the expressing role, which we numbered saying$_1$, saying$_2$, etc. It's tempting to think that, despite this, the saying relation is somehow more special than these other relations, in virtue of it being *the* saying relation. On the present view, this is just as much a mistake as thinking that Mt. Kilimanjaro is in some way privileged over the other mountain-like objects in its vicinity.

5 Applications to the Semantic Paradoxes

5.1 Semantic Pluralism and mysterious propositions

We have argued that when Alice utters, at t, the sentence 'everything Alice is saying at t is untrue' she says something, but not the proposition that everything Alice is saying at t is untrue. But if Alice didn't say this, a conspicuous question remains: what *did* she say? Let us now apply our considerations involving semantic pluralism to deflate this apparent mystery.

According to semantic pluralism, the word 'says' is associated with a bunch of distinct but equally important relations and so we have a whole bunch of candidate answers to the mysterious proposition worry. Perhaps, after Alice makes her utterance, she has said at t that everything she said* at t is untrue, where 'saying*' is another semantic relation that

[55] Throughout this discussion I've made quite a few assumptions here to ease readability. I've assumed, for example, that 'utterances' refer$_1$ and refer$_2$ to utterances, when it could be referring$_1$ and referring$_2$ to noises that are only slightly different from utterances. Similar assumptions were made about the other words used in my announcement.

plays the saying-role (but is for most purposes very similar to the saying relation).[56]

Notice, however, that given Semantic Pluralism the above discussion is overly simplistic: we also have a bunch of candidate questions too! One lesson we drew earlier was that there is nothing special about the saying relation: there's a question of what Alice said$_1$, what she said$_2$ and so on. Further instances of Prior's theorem allow us to get some idea of the situation (letting $i = 1, 2, 3, ...$)

> If Alice is saying$_i$ at t that everything she is saying$_i$ at t is untrue then at t she is saying$_i$ something true and something untrue.

Applying the same line of reasoning to each of the saying$_i$ relations, we might conclude that the proposition Alice is saying$_i$ at t, is the proposition that everything she is saying$_j$ at t is untrue for some j distinct from i.

We can predict these results by making assumptions about the meanings of the sub-sentential components of Alice's utterance. For example, given the assumption that 'says' means$_i$ the saying$_j$ in Alice's context, and some simplifying disquotational assumptions (that 'Alice' means$_i$ Alice etc) we can argue, by compositionality, that Alice's utterance of 'everything Alice is saying at t is untrue' says$_i$ that everything Alice is saying$_j$ at t is untrue.

Note, moreover, that the aforementioned compositionality assumption allows us to refute the following natural conjecture:

> 'says' means$_i$ the saying$_i$ relation.

For if it did then, given the same background assumptions, it would follow that at t Alice said$_i$ that everything Alice said$_i$ at t is untrue. But we rejected the latter hypothesis on the basis of Prior's theorem. Given the interconnected nature of expressing role, we must thus also reject the idea that 'means' means$_i$ the meaning$_i$ relation, 'true' means$_i$ truth$_i$, and other similar principles.

5.2 When is disquotational reasoning legitimate?

It is important to distinguish radical anti-disquotationalism, the thesis that most semantic relations do not relate 'Snow is white' to the proposition that snow is white, from the thesis that certain disquotational utterances are acceptable, such as the utterance "Snow is white' means that snow is white'. Conflating these two ideas often itself involves an instance of problematic disquotational reasoning.

RAD still leaves room for the acceptability of these disquotational utterances. For example, although most semantic relations do not relate 'snow is white' to the proposition that snow is white, it could still be true that some or all of these semantic relations relate utterances of the the disquotational sentence "snow is white' means that snow is white' to a true proposition. In which case, our general anti-disquotational sentiments notwithstanding, it could turn out that certain utterances stating disquotational principles will come out true in some or all the different senses of true. For example, a plausible sufficient condition for an utterance of this sentence to be true$_i$ would be if 'snow is white' meant$_i$ and meant$_j$ the same proposition, where meaning$_j$ is the relation 'means' means$_i$. More explicitly:

[56]Note that this idea generalizes to the mental versions of Prior's paradox. Prior's theorem entails, for example, that it's impossible to uniquely think that everything you are thinking is false. On my view when one produces a mentalese token that corresponds to the sentence 'everything I'm thinking is false', I do not think that everything I'm thinking is false, but a slightly different proposition: that everything I'm thinking* is false.

1. 'means' means_i meaning_j (this holds for some j according to the present view)

2. 'snow is white' means_i and means_j the same proposition: P.

3. "snow is white' means that snow is white' means_i that 'snow is white' means_j that P.

3 follows from 1 and 2 and compositionality given the simplifying assumption that the quotation name ' 'snow is white' ' refers_i to the sentence 'snow is white', and similar such assumptions. Given 2 'snow is white' means_j that P and so given 3, "snow is white' means that snow is white' means_i a truth, namely that 'snow is white' means_j that P.

Thus we have seen what it takes for a sentence of the form "P' means that P' to be true_i. Liar-like paradoxes will certainly place limitations on the extent to which even this form of disquotation can hold.[57] However, it's entirely possible to ensure that disquotational utterances of sentences that do not involve semantic vocabulary are acceptable. The main point to bear in mind is that the justification for why these utterances are acceptable, as we saw above, is not at all routine, and so we have at least a clearer picture of what happens when disquotational utterances involving paradoxical sentences fail to be true_i. In these cases, the meaning_i and meaning_j relations disagree about the proposition assigned to the paradoxical utterance (here meaning_j is what 'means' means_i).

Expanding on this a little, let us suppose for the sake of argument that semantic pluralism, and consequently our anti-disquotational conclusions, are indeed isolated to semantic words like 'says', 'means', and so on. Does it follow that we could continue to speak, more or less, as we are naïvely inclined to, only revising these practices in those highly unusual circumstances involving self-referential sentences containing semantic vocabulary? The present view predicts that naïve practice will lead us to assert falsehoods in much more mundane situations involving embedded speech reports. Suppose that Bob makes a speech report by uttering the sentence 'Alice said that 0=1'. I might in turn report on Bob by uttering the sentence 'Bob said that Alice said that 0=1'. Assume that 'says' means_1 saying_2 and means_2 saying_3, but that 'Alice', 'Bob' and '0=1' mean_1 and mean_2 Alice, Bob and that 0=1 respectively (in accordance with our assumption that semantic pluralism is isolated to semantic words). Then by compositionality for meaning_1, my utterance said_1 that Bob said_2 that Alice said_2 that 0=1. By compositionality for meaning_2, Bob said_2 that Alice said_3 that 0=1. Supposing Bob said_2 only one thing, and that the proposition that Alice said_2 that 0=1 is distinct from the proposition that Alice said_3 that 0=1, it follows that I said_1 a falsehood. For Bob said_2 that Alice said_3 that 0=1, he didn't say_2 that Alice said_2 that 0=1. One response is to go thoroughly error theoretic about embedded speech reports, but there are also more moderate responses. A particularly natural option is that 'says' is extremely context-sensitive to the extent that it can change it's meaning_1 within the utterance of a single sentence. In particular, in the sentence 'Bob said that Alice said that 0=1', the meaning_1 of 'says' at the beginning of the sentence is saying_2, but has shifted to saying_3 by the time we reach the second occurrence.[58]

5.3 Other versions of Prior's paradox

Let us examine again our variant of Prior's paradox, with an eye to bringing it under the umbrella of our proposed solution.

[57]For example, there are some instances of "S' is true if and only if S' that are classically inconsistent, and so assuming that the expressing relations are reasonably well-behaved no expressing relation can relate utterances of this sentence to truths.

[58]Related worries, and responses to them, are explored in a lot more detail in Dorr and Hawthorne [13].

Suppose, as before, that at t, Alice utters the sentence 'everything it sounded as though I said at t is false'. There are only two options consistent with Prior's theorem:

1. At t, it didn't sound as though Alice said that everything it sounded as though she said at t is false. (Corresponding to denying the antecedent of Prior's theorem: $S\forall p(Sp \rightarrow \neg p)$.)
2. There is at least one other proposition it sounded as though Alice said at t. (Corresponding to accepting the consequent: $\exists p(Sp \wedge p) \wedge \exists p(Sp \wedge \neg p)$.)

When put this way, it is evident that far from being a problem just for no-proposition theorists this paradox is a problem for everyone. Both of the options appear to be refuted by common sense: by uttering that sentence, it surely sounded as though Alice said that everything it sounded as though she said at t is false, and there wasn't anything else it sounded as though she said either. To say otherwise seems to contradict facts we appear to have direct introspective access to, namely how things sound to us.

But our remarks above put us in a position to better understand what is happening here. We have made the case that someone who has uttered the sentence 'snow is white' (say) has most likely not said that snow is white, but said some closely related proposition (since only a small proportion of the saying-like relations relate that utterance to the proposition that snow is white). Now let us consider for a minute the following plausible principle:

> In most circumstances what people have said and what it sounds as though they've said are the same.

I take it that this principle is a part of common sense. But in case it needs an argument, note that without it effective communication would be difficult. For it would mean that people would often sound as though they were saying something they weren't in fact saying, or conversely saying things without sounding as though they were saying them, neither of which is particularly conducive to communication.

What this means is that if disquotational principles are not a good guide to what people have said — having uttered the sentence 'snow is white' is not a good indicator that you have said that snow is white — then they are no better guide to what it sounds as though people have said. In this case, when Alice utters the sentence 'everything it sounded as though I said at t is false', she says some proposition P, but not the proposition that everything it sounded as though Alice said at t is false. Given our common sense principle, P is also the proposition it sounded as though she said.

Does this position cast doubt on the idea that we have introspective access to how things sound to us? I think not. I maintain that we know what Alice said, and what it sounded as though she said: P. Of course, disquotational reasoning is seductive, and if asked what it sounded as though Alice said, I might reply by uttering the sentence: 'it sounded as though Alice said that everything it sounded as though she said at t is false'. But this is not a failure of me to introspect on how things sound to me, but a failure to properly articulate in words the proposition it sounded as though Alice said. And it is not an unnatural mistake to make, given that P (the proposition Alice actually said) and the proposition I'd ascribe to Alice by naïvely applying a disquotational principle are closely related, and that I have the words readily available to make the latter speech report.[59]

[59]Note that if we adopt the intra-sentential-contextualist approach to the word 'says', indicated in the previous section, we have extra maneuvres available to us here. Suppose that, with her utterance, Alice said$_j$

The predicted mismatch between how things sound to us, and how we are naturally inclined to articulate how things sound to us, is arguably quite pervasive. To illustrate, consider an example involving context sensitivity.

Suppose that Alice and Bob are discussing their children, and Alice remarks on how much Bob's daughter, Enid, has grown: she utters the sentence 'Enid is tall'. In this context it seems to be perfectly fine to report what Alice said by uttering the sentence 'Alice said that Enid is tall'. But in other contexts it would be inappropriate. Suppose that Bob and Carol are attempting to get a book down from a high shelf, but neither are tall enough. Carol asks Bob whether Alice said anyone was tall, and he may correctly reply negatively. Indeed, if Enid came up as a potential book-reacher he could even say the following: 'No, Alice did not say that Enid is tall: she was talking about four-year-olds, and she said that Enid was tall-for-a-four-year-old'.

Indeed, I think we should go further and insist that it didn't even sound to Bob as though Alice said that Enid is tall. Given what we know about the context the utterance was made in, it sounds as though she was saying that Enid is tall-for-a-four-year-old — it's just obvious in this case that she wasn't saying that Enid is tall-for-an-adult, which is what 'Enid is tall' means relative to an ordinary context, or the context in which Bob and Caroline are talking, since it's clear to Caroline that Enid isn't tall in that sense (she cannot, for example, help them with their book). In accordance with our principle, what Alice said and what it sounded as though she said are the same: that Enid is tall for a child, or something similar.

But people can be bad at articulating how things sound to them. If you hear Alice utter the sentence 'Enid is tall' and you know it is a context different from your own, then you are typically in a position to know that she hasn't said that Enid is tall-for-your-context. However it is still very tempting to simply report Alice disquotationally, by saying 'Alice said that Enid is tall' or even 'it sounds to me as though Alice said that Enid is tall'. This might be tempting because the proposition Alice in fact said presumably cannot be described simply in your context because the word 'tall' doesn't express the appropriate property in your mouth. The proposition that Enid is tall, on the other hand, seems natural and simple because it's the proposition that I'm able to refer to in my context with a particularly simple definite description ('the proposition that Enid is tall'). Symmetrically from Alice's context that description would pick out her proposition, and it would be the proposition *I* picked out with it that would be hard for her to articulate. If Alice were to try to describe the proposition that Enid is tall in his context she would also find she didn't have simple words for it — the words 'the proposition that Enid is tall' would not succeed unless what 'tall' means in her context is the same as what it means in mine.[60]

that everything it sounded as though she said$_k$ is false. In accordance with our communicative principle, this is also what it sounded as though she said$_j$. I might be able to truthfully$_i$ utter the sentence 'it sounded as though Alice said that everything it sounded as though she said is false' because the word 'says' means$_i$ saying$_j$ in its first occurrence but saying$_k$ in its second occurrence (because there's been an intra-sentential context shift). Thus I say$_i$ a truth, namely that it sounded as though Alice said$_j$ that everything it sounded as though Alice said$_k$ is false.

[60]Other diagnoses of the above are possible too: if we don't know enough about what context Alice is in we won't know what she's said. In such cases there probably isn't a particular proposition you think she said, or a particular proposition which it sounds to you as though she said. Analogous moves may be made about paradoxical utterances: perhaps there's no particular proposition sounds as though Alice was saying. Or perhaps how things sound to me isn't very finely tuned, and there are several propositions it sounds as though she is or might be saying. In either case, Prior's theorem is upheld.

5.4 Hierarchies

Some ways of indexing the saying relations are more aesthetically pleasing than others. Since we have rejected the assumption that 'says' means_i saying_i, it's natural to choose an indexing in which whatever relation 'says' means_i is indexed by $i+1$. We can do this inductively by associating the saying relation with the index 0, and defining saying_{i+1} to be whatever relation 'says' means_i.

So indexed, the saying relations have the look of a hierarchy, and one might wonder how the resulting view compares to hierarchical approaches to the paradoxes.[61] There are some important differences. Firstly, not every expressing relation is arrived at via this process. It's plausible that there are uncountably many expressing relations, yet at most countably many expressing relations get indexed by this process (as there is no obvious way to continue our indexing conventions into the transfinite).[62] Conversely, we have no guarantee that every relation arrived in this way is an expressing relation. If meaning_i is a relation that barely satisfies the expressing role, it could be that 'says' means_i a relation that simply doesn't satisfy the expressing role.

Another significant difference is that for traditional hierarchical approaches to the paradox there is a clear direction to the hierarchy: one has a series of more expansive truth predicates, 'true_0', 'true_1', etc. where truth_{n+1} is supposed to capture truth for the language containing predicates 'true_0',..., 'true_n'. Thus the sentence 'something is true_1' is true_2, but 'something is true_2' is not true_1, as no sentence involving 'true_2' is true_1. There is no asymmetry of this sort in the present theory. Suppose we have established a language, as we apparently have, containing expressions for several different semantic concepts, 'meaning_1', 'meaning_2' et cetera. The principle of compositionality, for each possible index, dictates that every expression must have a meaning_1, a meaning_2 and so forth. Thus we may ask what the expression 'means_1' means_2 and 'means_2' means_1. It's clear that there is no constraint to the effect that expressions containing 'meaning_{n+1}' do not have meanings_n, however we have set up the indices. (The consistency of these sorts of compositionality principles is shown in the appendix.) It's also worth noting that we could have indexed our relations in the opposite order, i.e. by letting $n+1$ index the network of semantic concepts with the property that 'says' means_{n+1} saying_n.

On the other hand, there are also points of similarity with hierarchical views. For example, hierarchies of truth predicates typically come well-ordered by their indices, which among other things rules out loops: cases where one truth predicate appears below a another in the hierarchy, while the other also appears below the first. This raises a natural question about loops in the present framework. Could, for example, 'says' mean_1 saying_2 and mean_2 saying_1. Similar questions concerning loops of higher orders are also open: e.g. is it consistent that 'says' means_1 saying_2, means_2 saying_3 and means_3 saying_1?

Supposing we have indexed the saying relations in the way prescribed above, we can similarly rule out loops. This requires a generalization of Prior's paradox involving liar-like loops of arbitrary length, and is proved similarly. The theorem is stated in a language with n unary operators $S_1, ..., S_n$:

$$(\bigwedge_{1\leq i\leq n} S_n(\forall p(S_i p \to p)) \wedge S_{n+1}\forall p(S_1 p \to \neg p) \to \bigvee_{1\leq i\leq n+1} \exists p(S_i p \wedge p) \wedge \exists p(S_i p \wedge \neg p)$$

The theorem may be illustrated for the case of $n+1=3$, interpreting S_i as 'Alice said that', 'Bob said that' and 'Carol said that' for $i = 1, 2, 3$. The theorem states that if Alice

[61]See, e.g., Tarski [39], Burge [9] etc.

[62]An argument for the uncountability thesis, albeit in a slightly different context, is presented in Dorr and Hawthorne [13].

says that everything Bob says is true, Bob says everything Carol says is true, and Carol says that everything Alice says is false, then either Alice, Bob or Carol said a truth and a falsehood.

As with Prior's theorem, we may generate a puzzle by supposing that Alice utters the sentence 'everything Bob says is true', Bob utters 'everything Carol says is true' and Carol utters 'everything Alice says is false'. What does our proposed solution say about this case? Alice, Bob and Carol each simultaneously say$_1$, say$_2$ and say$_3$ something. Since 'says' means$_1$ saying$_2$ (given our indexing convention), we may conclude by compositionality that Alice says$_1$ that everything Bob says$_2$ is true. Similarly, we can infer that Bob says$_2$ that everything Carol says$_3$ is true, and Carol says$_3$ that everything Alice says$_4$ is false. But suppose we had a loop of order three: that saying$_1$=saying$_4$. Then we have that Carol says$_3$ that everything Alice says$_1$ is false. Now we may apply our generalization of Prior's theorem, reading S_1 as 'Alice said$_1$ that', S_2 as 'Bob said$_2$ that' and S_3 as 'Carol said$_3$ that'. We may conclude that either Alice said$_1$ a truth and a falsehood, Bob said$_2$ a truth and a falsehood or Carol said$_3$ a truth and a falsehood; each of these options are inconsistent with our proposed constraints on saying relations.

5.5 Similarity and disquotational reasoning

In this section we'll elaborate a bit further on the idea that the expressing relations are similar to one another and thereby relate sentences to similar propositions. We'll explain how this is relevant to the apparent acceptability of disquotational reasoning when applied to extensional semantic concepts like truth.

Our discussion in section 4.2 suggested that 'snow is white' likely does not mean that snow is white, on the grounds that very few expressing relations, of which *saying* is just one, relate 'snow is white' to this proposition. But these expressing relations collectively relate 'snow is white' to a cluster of closely related propositions. Thus one might expect the following to hold:

> 'snow is white' means a proposition close to the proposition that snow is white.

Here we will resume our earlier formal model, defining two propositions to be *close* if there is a perturbation that maps one to the other. There is, of course, nothing special about the meaning relation, so we might also wish to investigate the claim that the propositions 'snow is white' means$_1$, means$_2$, means$_3$ and so on, are all close to the proposition that snow is white.

Let us investigate some consequences of this idea. Firstly, it offers an explanation for the apparent acceptability of many instances of the oft-discussed T-schema. Once one has accepted the possibility that 'snow is white' does not mean that snow is white, we have no logical guarantee that the proposition 'snow is white' means holds if and only if snow is white. Assuming our account of sentential truth as meaning a true proposition, the upshot is that we have no guarantee that:

> 'snow is white' is true if and only if snow is white.

Nonetheless, it seems extremely plausible. We might substantiate this as follows: 'snow is white' means some proposition, P, which is close to the proposition that snow is white. Since snow is white couldn't easily have been false — it is true in worlds sufficiently like our own — any proposition close to snow is white must also be true, and so P must true, and thus the sentence 'snow is white' is true. Again, there is nothing special about truth, and the preceding considerations also generalize to suggest that 'snow is white' is true$_i$ if and only if snow is white, for each i, given the parallel assumption that 'snow is white' means$_i$ a proposition close to the proposition that snow is white.

The above argument rested on the idea that if P and Q are similar, and P is true in all nearby worlds, then Q is also true. This is a consequence of our model of propositional similarity; I put the argument in a footnote.[63]

Exceptions to the T-schema arise only when the propositions involved are false in nearby worlds. Lets try and make this a bit more precise. Say that a proposition is determinate iff it's true at all worlds sufficiently similar to actuality. It may be given a simple Kripke-style semantics as follows:

Determinacy It's determinate that P at a world w iff P is true at every world v such that $d(w, v) \leq \alpha$

Propositional determinacy and similarity are related as follows: A proposition is determinate if and only if every proposition close to it is true.[64] Say that it's determinate whether P when either P or its negation is determinate. Then we may use the assumption that 'P' means something close to P to establish:[65]

RT If it's determinate whether P then 'P' is true if and only if P.

I have argued elsewhere that a principle of this form is the correct way to capture the restriction on disquotational reasoning about truth (Bacon [2]).

Thus we may make some attractive predictions from the assumption that 'snow is white' means something close to the proposition that snow is white. But given the misgivings we have expressed so far about disquotational reasoning, it's natural to ask why we should treat this disquotational thought any differently? More importantly, what about a more general disquotational *schema*, of the form:

Closeness schema 'P' means something close to the proposition that P.

It's not obvious what it means to simply 'endorse' a schema like this. One might take it as endorsing an infinite number of propositions, that includes things like the claim that 'snow is white' means something close to the proposition that snow is white, that 'grass is green' means something close to the proposition that grass is green, *and so on, and so forth.* But it is surprisingly hard to spell out exactly what 'and so on, and so forth' amounts to here. We need to figure out which pairs of sentences and propositions can be plugged into the left and right, respectively, of this principle. But the question of whether there is a disquotation operation, mapping sentences to propositions, is ill-posed; talk of disquotation is at best an

[63]Suppose that w is the actual world P is true at every nearby world: for every x such that $d(w, x) \leq \alpha$, $x \in P$. Suppose also that Q is close to P. Then, without loss of generality, there is some perturbation, i, such that $iP = Q$. Since i is perturbation, $d(w, iw) \leq \alpha$, and since P is true in all worlds x such that $d(w, x) \leq \alpha$, $iw \in P$. Since $iQ = \{ix \mid x \in Q\} = P$, it follows that w the actual world, is in Q so Q is true as required.

[64]If P is determinate, then for any perturbation, i, iP is true: if w is the actual world, $i^{-1}w$ is close to actuality (since i^{-1} is also a perturbation), and so $i^{-1}w \in P$, and thus $w \in iP$, so iP is true. Conversely, if every perturbation of P is true, and v is close to w (actuality), then the mapping i which switches w and v and leaves everything else alone is a perturbation. So iP is true, i.e. $w \in iP$, and thus $v = i^{-1}w \in P$. More details can be found in Bacon [4] chapter 13, which investigates structurally similar analyses of determinacy in the context of vagueness.

[65]The case where P is determinate was argued for above. In case $\neg P$ is determinate we need another lemma, which also falls out of our account of propositional similarity: namely that if P is false in all nearby worlds, and Q is close to P then Q also false.

abbreviation for something metalinguistic. Alternatively, one might endorse the claim that every sentence of the form "P' means something close to the proposition that P' is true. Here it is clear what it means to be a sentence of this form: one simply must replace the letter 'P' with the same sentence in both of its occurrences.

Interestingly, we can prove the truth of all instances of the closeness schema from a few natural, and for the most part non-disquotational, assumptions, which I offer in an exploratory spirit. This gives some insight into what it takes for this schema to have only true instances. Here are the assumptions:

1. The relation that 'means' means is similar to meaning.

2. 'close' expresses a relation that holds between propositions that bear similar expressing relations to 'P'.

3. "P" means 'P', 'something' means something.

Suppose that 'P' means Q, 'means' means meaning$'$, and 'close' means close$'$. By compositionality and 3 we have that "P' means something close to P' means that 'P' means$'$ something close$'$ to Q. By 1, means and means$'$ are similar relations. Thus they relate the sentence 'P' to close$'$ propositions by 2. Since 'P' means Q, it follows that 'P' means$'$ something close$'$ to Q. Since this is what "P' means something close to P' means, and we have just established it to be a true proposition, it follows that "P' means something close to P' is true.

Similar arguments may be used to derive the conclusion that "P' means something close to P' is true$_i$ for each i, provided we make assumptions parallel to 1-3 for meaning$_i$.

As with all of the disquotational principles we have considered, there is the threat of paradox. For the closeness schema, let C be the sentence 'something close to Cs meaning is false', and consider an instance of Prior's theorem interpreting S as 'C means something close to the proposition that'.

> If C means something close to the proposition that something close to Cs meaning is false, then there's a truth and a falsehood close to Cs meaning.

Given than C='something close to Cs meaning is false', the antecedent is an instance of the closeness schema. So we may infer the consequent, and in particular, infer that something close to Cs meaning is false. We might see this as another version of the practical bind we found the no-proposition theorist in: for we have just proven that something close to Cs meaning is false. In writing this conclusion, I have written an instance of C itself. And so I have thereby asserted a proposition that is close to a false proposition (since something close to Cs meaning is false). Or, equivalently, I have asserted a proposition that is not determinate.

Determinacy is plausibly a necessary condition for knowledge, and thus proper assertion. An indeterminate proposition is one which is false in a nearby world, and a so belief in an indeterminate proposition runs the risk of being unsafe, and thus unknown. Since uttering C assertively results in one asserting an indeterminate proposition, we should refrain from uttering C, or sentences that entail C. *In particular, we should refrain from assertively uttering some instances of the closeness schema.* This is not necessarily because these instances express falsehoods (we provided a tentative argument that they express truths above), but because they express truths that could easily be false.

6 Conclusion

It is typical, in presentations of the liar paradox, to frame the paradoxes as exceptions to a class of normally acceptable principles.[66] According to the standard perspective, the disquotational principles tell us important and non-accidental truths about the relation of language to the world: that 'Mt. Kilimanjaro' refers to Mt. Kilimanjaro, that 'Snow is white' means that snow is white, and so on and so forth. What the paradoxes tell us, according to this outlook, is that while these principles are usually true, and indeed encapsulate something important when true, they admit of exceptions distinctive to sentences that involve semantic notions.

Once this outlook is accepted, it becomes all but inevitable to see the philosopher's role, in providing a solution to the paradoxes, to be that of characterizing the exceptions to the disquotational principles: to offer some general theory of what those problematic instances look like, and maybe offer some explanation for why sentences of that sort 'go wrong'. Attempts to solve the paradoxes based on this premise have inevitably led to revenge paradoxes: for one can always ask about the sentence that classifies itself as either untrue or as one of the exceptions, according to whatever analysis one offers of 'exception', and one finds oneself in paradox once again.[67]

The attitude of the present paper is that it is a mistake in the first place to frame the paradoxes in terms of exceptions, and things 'going wrong'. Disquotational principles do not in general express truths, even when they do not involve semantic words, and when instances of the principles do express truths it is as best an accident, as these instances will typically bear other equally important semantic relations to falsehoods. Rather than characterise the exceptions to disquotational reasoning, the challenge, on the present view, instead becomes to explain why disquotational principles ever true express truths — a task that will presumably require us to investigate how speakers of a language use both semantic and non-semantic words (see the exploratory remarks in the previous section). I submit that this is entirely appropriate: if it is a fact that 'snow is white' means that snow is white in English, it is a contingent and empirical matter, and it is a virtue that our approach predicts this.

[66] I include my former self here; see Bacon [2].

[67] In other work (Bacon [2]) I have spelled out what I take to be the general form of this paradox for the classical logician. There I also offered an alternative revenge-free approach that still falls within the paradigm of trying to offer an analysis of the 'exceptions' to the disquotational principles.

A Formal theories of expression

In this section we develop a formal framework for representing the vision of semantics we've developed in the previous sections. We'll work our way up to this in stages. First, we'll present a formal theory of meaning that allows us to theorize compositionally about the semantic values of arbitrary expressions. This theory is superficially semantically monist, in that concerns a single notion of meaning. Second, we'll see that the theory proves 'meaning' doesn't mean meaning, and we investigate a sequence of stronger theories, the first of which states that the relation 'meaning' means also satisfies the formal axioms of a meaning function, suggestive of a view in which there are multiple meaning relations. Third, we look at languages that are a bit more explicitly friendly to semantic pluralism, containing the predicate of *being an expressing relation*. Finally, we'll examine the idea that the relevant meaning relations are 'close' to one another, in some sense to be specified, and use this fact to predict the truth of certain disquotational principles.

Formal work on the paradoxes has focused on theories of truth. For anyone interested in the wider implications of the paradoxes for semantics, this fairly narrow focus is unfortunate. Firstly, truth is a semantic property of sentences, and so doesn't say anything about the semantic properties of subsentential expressions, such as names, predicates, operators, quantifiers and so forth. Secondly, truth is an extensional notion: two sentences might have the same truth value but express different propositions. Thus we are missing out on important semantic properties of even sentences.[68]

In order to broaden our aspirations in the first direction we need a language suitable for theorizing about the meanings of expressions with arbitrary grammatical types. As we suggested in section 2.1 the natural setting for this is type theory. In order to accommodate the second issue, we must choose a version of type theory that does not include the principle that there are only two propositions: the true and the false.

Each expression of our language will have a unique type. There is the type of singular terms, e, the type of sentences, t, and for any types σ and τ, there is the type of expressions that combine with argument expressions of type σ to produce an expression of type τ: $\sigma \to \tau$. We omit outer parenthesis and parenthesis that are associated to the right. Thus, for example, $t \to t \to t$ is short for $(t \to (t \to t))$: an expression which takes two sentences in succession to produce a sentence (a binary connective). Our language will contain the following primitive expressions (constants) of the following types:

1. $\wedge$ of type $t \to t \to t$, $\neg$ of type $t \to t$
2. $\forall_\sigma$ of type $(\sigma \to t) \to t$ for each type σ
3. $[\![\cdot]\!]^\sigma$ of type $e \to \sigma$
4. An infinite stock of *expression* constants, $c_1, c_2, ...$ of type e.

According to our explanations of the types $\neg$ combines with a sentence to form a sentence, i.e. it is an operator. The type of $\wedge$ is that of a binary connective, as explained above. $\forall_e$ takes a predicate — say, F of type $e \to t$ — and produce a sentence $\forall_e F$, intuitively

[68] McGee [29] chapter 1 shows how certain paradoxes of reference and of extension can be simulated within a formal theory of satisfaction (a theory which tells us how to evaluated open sentences with respect to variable assignments). But the simulated notions are still extensional (e.g. we may talk about a predicates extension, but not the property it expresses). Moreover, we seek a theory that treats the semantic properties of arbitrary expressions in the type hierarchy.

stating that every individual is F. More generally, $\forall_\sigma$ combines with a predicate of type σ, F of type $\sigma \to t$, to create a higher-order quantified sentence $\forall_\sigma F$. The constants $c_1, c_2, ...$ on their intended interpretations will denote expressions of the language; we call the set of expression constants E. $[\![\cdot]\!]^\sigma$ has c_k as an argument, which denotes an expression of α the language, and the result $[\![c_k]\!]^\sigma$ is an expression of type σ, supposed to be the meaning of α.

For each type σ, we help ourselves to an infinite stock of variables, Var^σ of that type. We form complex expressions as follows

- Every constant described above is an expression of type σ
- Every variable of type σ is an expression of type σ
- If α is an expression of type $\sigma \to \tau$ and β an expression of type σ then $(\alpha\beta)$ is an expression of type τ
- If α is an expression of type τ and $x \in Var^\sigma$, $\lambda x\, \alpha$ is an expression of type $\sigma \to \tau$

If ϕ is a sentence (an expression of type t) and x a variable of type σ then we follow the usual convention of writing $\forall_\sigma x\phi$ as short for $\forall_\sigma \lambda x\, \phi$.

We may introduce other logical expressions, such as $\vee$, $\to$ and $\exists$, by definition. An important example is a generalized notion of identity $=_\sigma$ of type $\sigma \to \sigma \to t$, which takes two expressions of type σ, a and b, and produces a sentence that informally states the identity of a and b. It may be defined as $\lambda x\lambda y\forall_{\sigma\to t}X(Xx \to Xy)$. When it is clear from context, we omit the type subscript from the identity relation.

Two terms, α and β, are called $\beta\eta$-equivalent if one can be obtained from the other by some sequence of the following operations: (i) renaming bound variables with new variables, (ii) replacing a subterm of the form $(\lambda x\alpha')\beta'$ with $\alpha'[\beta'/x]$ where β' is substitutable for x in α', or vice versa, (iii) replacing a subterm of the form $\lambda x\, \alpha' x$ with α', where x doesn't appear free in α, or vice versa. Two particular families of expressions are important. For each σ and τ there is an expression K of type $\sigma \to \tau \to \sigma$ defined as $\lambda x\lambda y\, x$ where $x \in Var^\sigma, y \in Var^\tau$. Similarly, for any ρ, σ and τ there is an expression S of type $(\rho \to \sigma \to \tau) \to (\rho \to \sigma) \to \rho \to \tau$ defined as $\lambda x\lambda y\lambda z\, xz(yz)$. It is a well-known fact that every closed term is $\beta\eta$-equivalent to a term built out of combinations of S, K and the constants via application. This fact proves useful because it allows us to provide a compositional axiomatization of expression without worrying about variables, variable assignments, and so on. In addition to S and K, we also adopt the abreviation I for $\lambda x\, x$.

Finally, note that the official syntax uses prefix conventions. However, we will use use infix notation for the connectives, writing e.g. $A \wedge B$ and $A =_t B$ instead of $(\wedge A)B$ and $(=_t A)B$. And we will use 'outfix' notation for the semantic primitives, writing $[\![c]\!]^\sigma$ instead of $[\![\cdot]\!]^\sigma c$.

A.1 Representing syntax

In order to theorize about the semantics of expressions we need to be able to theorize about expressions themselves. Expressions are individuals, and so we have supposed that we have an infinite supply of type e constants, each to be thought of naming a different expression of our language. Since we are just concerned with modeling various sorts of self-referential sentences, we will work with an extremely pared down theory of syntax. (A more expressive theory of syntax would desirable if we were attempting to do anything more serious than provide a proof of concept. In formal work on truth, this is often achieved by representing syntax indirectly by encoding it in arithmetic; I'll offer some more direct ways of doing this below.)

For each expression, α, of our language, we assume there is some individual constant c_k denoting that expression. Obviously it doesn't matter which constant plays this role, so long as no two expressions get associated with the same constant; i.e. the assocation must be injective. We shall call such an association a *representation* of the syntax.

Definition 1. *A representation function is an injective function* $rep : \mathcal{L} \to E$.

An example from English is instructive. Given an expression of some type, e.g. *loves* of $e \to e \to t$, there is an expression of type e, *'loves'*, which is the quotation name for the expression *loves*. Note that distinct expressions always get mapped to distinct names, as can be seen by noting the mapping from expressions to their underlying strings is injective. For a slightly more formal example:

Example 1 (Gödel quotes)**.** *Suppose $\mathcal{L}$ contains the language of arithmetic, and we have a Gödel numbering of the expressions of $\mathcal{L}$. Then the mapping which maps α to the numeral of $\ulcorner \alpha \urcorner$ is injective and thus a representation function.*

The difference between numerals and numbers won't play much role in the following. Thus we shall often use Gödel numbering style notations, such as $\ulcorner \cdot \urcorner$ and $\langle \cdot \rangle$, for discussing specific examples of representation functions, but the more familiar prefix notation *rep* when making generalizations about them. We will also adopt a convention about embedded uses of representations. For example, if $\langle \alpha \rangle = c_7$, then we will write $\langle F \langle \alpha \rangle \rangle$ for $\langle F c_7 \rangle$. (Note that the string of symbols $F\langle \alpha \rangle$ does not correspond to an expression of the language, and so technically is not in the domain of $\langle \cdot \rangle$, whereas Fc_7, by contrast, is. Thus this convention is not automatic, but imposed.) We'll similarly abuse notation by using our abbreviations inside the representation function, e.g. writing $\langle K \rangle$ for what really should be $\langle \lambda x \lambda y\, x \rangle$.

Every expression gets associated with an expression constant. Because it doesn't matter which constant represents which expression, there are lots of ways to do this. But we also see that not all representations are 'isomorphic', since we have to choose where to send expressions that themselves contain expression constants. For example, suppose F is a predicate of type $e \to t$, then some representation functions have self-referentiality by mapping $F(c_1)$ to c_1, where as others might not, by e.g. mapping $F(c_k)$ to a distinct expression constant for every c_k. Note for example, that the example of quotation doesn't generate self-reference: one way to see this by composing the quotation representation function with the function from expressions to the length of their underlying strings of symbols: since quotation always increases the length by two, we see that if c is a quotation name, the length of $F(c)$ is always greater and is thus always mapped to something distinct from c by quotation.

We write $e^n \to t$ as short for the type of n-ary predicates, i.e. $e \to ... \to e \to t$. This notational convention generalizes: in the next definition we write $\sigma^0 \to \tau$ for τ and $\sigma^{n+1} \to \tau$ for $\sigma \to \sigma^n \to \tau$

Definition 2. *Suppose that $\phi_1, ...\phi_n$ are expressions of types $e^n \to \sigma_1, ..., e^n \to \sigma_n$. A fixed point, relative to a representation function rep, is a sequence of distinct individual constants $c_1, ..., c_n$ such that:*

- *$rep(\phi_i c_1 ... c_n) = c_i$ when $1 \leq i \leq n$.*

The reader may have noticed that our notion of a fixed-point is bit more general than the usual one. Firstly, it is a conception of fixed-points for n-ary functions. The familiar liar sentence arises when $rep(\neg Tc) = c$ for some constant c, assuming T is a truth predicate, and arises from our more general conception of a fixed point when $n = 1$ and $\phi_1 = \neg T$. Secondly, even in the unary case, we can ask for fixed-points for arbitrary expressions of type $e \to \sigma$, in contrast to the familiar fixed-points of predicates (of type $e \to t$). For example, the expression $\lambda x \lambda p(\neg([\![x]\!]^{t\to t}p))$ has type $e \to (t \to t)$. A fixed-point of this expression is a constant c such that $c = \langle \lambda x \lambda p(\neg([\![x]\!]^{t\to t}p))\rangle$. Informally, $c =$ 'the result of composing the meaning of c with negation'. It is easily seen that any sequence of expressions $\phi_1 ... \phi_n$ of the indicated types have fixed-points relative to some representation function, by simply stipulating that $rep(\phi_i c_1 ... c_n) := c_i$ for each i, and extending rep to any injective function on the remaining expressions.

A substitution of the expression constants is a function $i : E \to E$, mapping expression constants to expression constants. Given an arbitrary expression α of type σ, we may *apply* a substitution i to α to produce another term of type σ, $i\alpha$. $i\alpha$ is defined as the result of replacing each expression constant c_k appearing in α with $i(c_k)$.

Definition 3 (Translation of representations)**.** *A translation from one representation function rep, to another rep$'$, is a substitution $i : E \to E$ such that $rep'(i\alpha) = irep(\alpha)$ for every expression α.*

$$\begin{tikzcd} \mathcal{L} \arrow[r, "rep'"] & E \\ \mathcal{L} \arrow[u, "i"] \arrow[r, "rep"] & E \arrow[u, "i"'] \end{tikzcd}$$

If i is a translation from rep to rep$'$, we write $i : rep \to rep'$.

To illustrate the general idea, suppose that $\ulcorner \cdot \urcorner$ and $\langle \cdot \rangle$ are two representation functions, and i a translation from the former to the latter. Consider first an expression that does not itself contain any expressions constants, such as K of type $e \to t \to e$. $\ulcorner K \urcorner$ and $\langle K \rangle$ are thus two expression constants, but because they are different representations they may not be the same expression constant. The translation i simply tells us how to move between one and the other: in particular we get that $i : \ulcorner K \urcorner \mapsto \langle K \rangle$. This follows because $iK = K$, since K doesn't contain any expression constants, and because $i\ulcorner K \urcorner = \langle iK \rangle$ by the definition of a translation. What happens if we apply i to an expression containing expressing constants itself, such as $\ulcorner K \ulcorner K \urcorner \urcorner$? By applying the definition we get that

$$\ulcorner K \ulcorner K \urcorner \urcorner \mapsto \langle K \langle K \rangle \rangle$$

Firstly, $i\ulcorner K \ulcorner K \urcorner \urcorner = \langle i \ulcorner K \urcorner \rangle$, and we also just showed that $i \ulcorner K \urcorner = \langle K \rangle$. We may similarly see that $\ulcorner K \ulcorner K \ulcorner K \urcorner \urcorner \urcorner \mapsto \langle K \langle K \langle K \rangle \rangle \rangle$, and so on. Informally, we may think of the commuting square condition on a translation as ensuring that we replace $\ulcorner \cdot \urcorner$ everywhere with $\langle \cdot \rangle$, *including the cases where they appear embedded.*

Definition 4. *A representation $rep : \mathcal{L} \to E$ is*

1. *Universal iff for every representation function $rep' : \mathcal{L} \to E$ there is at least one translation $i : rep' \to rep$.*

2. *Terminal iff that translation is unique.*

3. *Couniversal for every representation function $rep' : \mathcal{L} \to E$ there is at least one translation $i : rep \to rep'$.*

4. *Initial iff that translation is unique*

We shall be primarily interested in universal representation functions because, as we shall argue below, they instantiate every sort of self-reference: every sequence of expressions has a fixed-point. In a terminal representation function these fixed points are moreover unique.

The difference between couniversal and initial representation functions will be less important for us: these representation functions contain no self-reference. Informally here is how one might construct an initial representation function. Firstly, consider the language fragment that contains no expression constants, call it $\mathcal{L}_0$, and then proceed inductively, constructing $\mathcal{L}_{n+1}$ as the result of adding a new expression constant for each expression of $\mathcal{L}_n$, making sure to leave infinitely many expression constants unused at each stage (so there are infinitely many expression constants left for the next round). One reaches a fixed point of this process at $\mathcal{L}_\omega$, defined as the union of each $\mathcal{L}_n$. We may think of this as a precise version of the informal example of quotation in English we introduced earlier.

Intuitively a universal representation function has every sort of self-reference. This follows from a more general fact. If $i : rep' \to rep$, and if $\phi_1, ..., \phi_n$ have fixed points under the representation rep', then $i\phi_1, ..., i\phi_n$ have fixed points under rep. Specifically, if $c_1...c_n$ are the fixed points under rep' then $ic_1, ..., ic_n$ are the fixed points under rep. We can illustrate with the liar sentence: Suppose that we had a truth predicate, T, in the language and that c is a liar sentence with respect to rep': $rep'(\neg Tc) = c$. Then $rep(\neg T(ic)) = rep(i\neg Tc) = irep'(\neg Tc) = ic$, so ic is a fixed point of the predicate $\neg T$ with respect to rep. In other words, if liar-like self-reference is possible at all (i.e. relative to any representation function) then liar-like self-reference occurs in a universal representation function. Universal and terminal representation functions exist, although the details will take us to far afield.

A.2 The compositional theory of expression

The following theory, and the subsequent consistency proofs, are formulated relative to some fixed representation function $\langle\cdot\rangle$. So that we may freely help ourselves to the existence of particular paradoxes, we will assume that $\langle\cdot\rangle$ is universal, although nothing of substance turns on this choice. We begin by outlining a basic theory of expression encoding the assumption that the $[\![\cdot]\!]$ functions behave compositionally.

Compositionality

C(app) $[\![\langle\alpha\beta\rangle]\!]^\tau = [\![\langle\alpha\rangle]\!]^{\sigma\to\tau}[\![\langle\beta\rangle]\!]^\sigma$

C($\forall$) $[\![\langle\forall\rangle]\!]^{(\sigma\to t)\to t} = \forall$

C($\wedge$) $[\![\langle\wedge\rangle]\!]^{t\to t\to t} = \wedge$

C($\neg$) $[\![\langle\neg\rangle]\!]^{t\to t} = \neg$

C(S) $[\![\langle S\rangle]\!]^{(\sigma\to\tau\to\rho)\to(\sigma\to\tau)\to\sigma\to\rho} = S$

C(K) $[\![\langle K\rangle]\!]^{\sigma\to\tau\to\sigma} = K$

C(exp) $[\![\langle\langle\alpha\rangle\rangle]\!]^e = \langle\alpha\rangle$

$[\![\beta\eta]\!]$ $[\![\langle\alpha\rangle]\!] = [\![\langle\beta\rangle]\!]$ when α and β are $\beta\eta$ equivalent.

Apart from the axioms governing semantic notions, we must also ensure that expressions are distinct:

Exp $\langle\alpha\rangle \neq \langle\beta\rangle$

where α and β are distinct expressions. We name CE the result of combining these axioms with a standard axiomatisation of higher-order logic, H.[69]

CE is little more than a showcase of a system capable of semantic theorizing: there are several ways in which our theory may be strengthened and improved. Firstly, our theory does not have a fully fleshed out theory of syntax. A more elaborate theory might include terms that internalize important syntactic operations, such as a term, *app*, of type $e \to e \to e$ subject to the law $app\,\langle\alpha\rangle\,\langle\beta\rangle = \langle\alpha\beta\rangle$, and T^ρ of type $e \to t$ stating that something is an term of type ρ, subject to the schema $T^\rho(\langle\alpha\rangle)$ when α has type ρ and $\neg T^\rho(\langle\alpha\rangle)$ when α doesn't.[70]

Second, with a richer theory of syntax, some of the schematic principles, such as C(app), can be replaced by the stronger quantified principles. Every instance of C(app), for example, may be deduced from the following quantified principle $\forall_e xy(T^{\sigma\to\tau}x \wedge T^\sigma y \to [\![app\,x\,y]\!]^\tau = [\![x]\!]^{\sigma\to\tau}[\![y]\!]^\sigma)$. We won't spell these stronger theories out explicitly, but the constructions to follow easily generalize to show the consistency of these stronger theories.

Third, we have taken as primitive a semantic expressing function that maps closed expressions to denotations, and have appealed to the functional completeness of the S and K combinators to deal with the jobs usually done with variables and variable assignments. This greatly simplifies the exposition, but one might argue that it doesn't properly reflect actual semantic theorizing which usually favors reasoning with variable assignments.[71] In order to develop a richer theory that can talk about the meanings of open expressions relative to assignemnts to the variables, we must instead take as primitive an expressions $[\![\cdot]\!]^{\sigma_1\ldots\sigma_n\tau}$ of type $e \to e^n \to \sigma_1 \to \ldots \to \sigma_n \to \tau$, which should be thought of as taking a type e entity representing an open expression α of type τ, n type e entities representing variables $x_1 \ldots x_n$ of types $\sigma_1, \ldots, \sigma_n$, things $a_1 \ldots a_n$ of types $\sigma_1, \ldots, \sigma_n$, and results in something of type τ to be thought of as the semantic value of α, when x_1 is interpreted by a_1, x_2 by a_2 and so on (and produces some default value if α or its free variables do not have the right type).

Fourth, and finally, one has to do some computation to figure out what the instances of the schematic principle $[\![\beta\eta]\!]$ are (i.e. we have to first prove that α and β are $\beta\eta$-equivalent before we know that they maybe inserted in the principle). It is in fact possible to reduce this unwieldy principle to a finite collection of type-schematic equations, called the Curry equations.[72] These equations, which include principles like $[\![S(KI)]\!]^{\sigma\to\sigma} = [\![I]\!]^{\sigma\to\sigma}$ (suitably typed), wear their logical form on their sleeve (although each equation still has infinitely many instances, due to the fact there are infinitely many varieties of S, K and I depending

[69] H consists of analogues of the standard axioms of first-order logic for the higher-order quantifiers, and the principle that $\beta\eta$-equivalent terms of substitutable. For a concrete axiomatization, which I'll assume in what follows, see Bacon [3].

[70] In principle any operation on the syntax can be internalized. A noteworthy example would be the representation function $\langle\cdot\rangle$ itself: this would be a term f of type $e \to e$ subject to the schematic principle $f(\langle\alpha\rangle) = \langle\langle\alpha\rangle\rangle$.

[71] Although, following Jacobson [27], there is a notable tradition among linguists of doing semantics in a variable-free setting.

[72] See Hindley and Seldin [25], p86.

on how they are typed). An alternative approach is to adopt the variant system mentioned above for theorizing about open expressions and variable assignments. A paradigm example of semantic reasoning about variables and variable assignments is the proof that the semantic values of $\beta\eta$-equivalent terms are in fact identical. Thus one should expect to actually be able to prove the principle $[\![\beta\eta]\!]$ within a suitable theory of semantic expression with variable assignments, and a suitably rich theory of syntax.

Here is a simple theorem in the above theory:

$$[\![\langle\neg A\rangle]\!] = \neg[\![\langle A\rangle]\!].$$

It is proved by noting that the right-hand-side is identical to $[\![\langle\neg\rangle]\!][\![\langle A\rangle]\!]$ by C(app), and that $[\![\langle\neg\rangle]\!] = \neg$ by C($\neg$).

Some non-theorems are also worth remarking upon. For example, consider the disquotational schema:

$$[\![\langle\alpha\rangle]\!]^\sigma = \alpha$$

where σ can be replaced by any type, and α by any expression of that type. Consider a fixed point $\langle\neg[\![c]\!]^t\rangle = c$ (which exists by the universality of $\langle\cdot\rangle$). By the fixed-point identity, $[\![c]\!]^t = [\![\langle\neg[\![c]\!]^t\rangle]\!]$, and from the disquotational schema we get $[\![\langle\neg[\![c]\!]^t\rangle]\!] = \neg[\![c]\!]^t$, from which we could conclude the inconsistent $[\![c]\!]^t = \neg[\![c]\!]^t$. This instance of the disquotational schema is thus not a theorem unless the theory is inconsistent (we prove the consistency of the theory shortly). This observation also allows us to derive our earlier claim that 'means' doesn't mean meaning. One instance of this idea is the following:

$$[\![\langle[\![\cdot]\!]^t\rangle]\!]^{e\to t} \neq [\![\cdot]\!]^t$$

For using compositionality we may decompose $[\![\langle\neg[\![c]\!]^t\rangle]\!]$ as follows:

$$[\![\langle\neg\rangle]\!]^{t\to t}([\![\langle[\![\cdot]\!]^t\rangle]\!]^{e\to t}([\![c]\!]^t))$$

(i.e. the meaning of negation applied to the meaning of 'means' applied to the meaning of c). We know that $[\![\neg]\!]^{t\to t} = \neg$ by C($\neg$), $[\![c]\!]^e = c$ by C(exp), so given the above identity we could conclude that the above was the same as $\neg([\![\cdot]\!]^t c)$, or applying our notational conventions, $\neg[\![c]\!]^t$. This is exactly the assumption that we just observed to lead to contradiction.

Here is another non-theorem (the fact that it is a non-theorem will become evident when we start looking at models):

$$[\![\langle[\![\langle\neg\rangle]\!]^{t\to t} = \neg\rangle]\!]^t = \top$$

It says that the sentence C($\neg$) expresses the tautologous proposition. We could in fact derive this identity, via compositionality, if we had the assumption that $[\![\langle[\![\neg]\!]^{t\to t}\rangle]\!]^{t\to t} = [\![\langle\neg\rangle]\!]^{t\to t}$ (as CE already proves $[\![\langle=_t\rangle]\!]^{t\to t\to t} =_{t\to t\to t} =_t$ and $[\![\langle\top\rangle]\!] = \top$, so the above would amount to the easily proven $([\![\langle\neg\rangle]\!]^{t\to t} = [\![\langle\neg\rangle]\!]^{t\to t}) = \top$). CE proves that $[\![\langle\neg\rangle]\!]^{t\to t} = \neg$, so if the above fails it must be because $[\![\langle[\![\neg]\!]^{t\to t}\rangle]\!]^{t\to t} = \neg$ is false. Thus, if there are failures of the above principle, then whatever 'means' means, it does not play the expressing role, since does not map 'negation' to negation.[73] This leads us to consider natural strengthenings of CE which we'll consider next.

[73]The expressing role involves certain linguistic practices: specifically connections between the utterance of sentences involving negation to the saying of propositions. The meaning of negation in a language like English, say, is pretty much entrenched to the extent that it seems unlikely there would be a compositional collection of expressing relations in which negation did not denote the negation operation.

A.3 Semantic Pluralism

The specter of semantic pluralism arises even in our simple language containing a single family of expressing predicates. There's a natural extension of the theory CE that states that the relation that 'means' means behaves formally like a meaning function, in that it is also governed by the axioms CE. Let us call CE^1 the theory you get by replacing each axiom of CE with the corresponding claim with two iterations of $[\![\cdot]\!]$, and combining it with CE. It is helpful to have a shorthand for this: we will write $[\![\cdot]\!]^\sigma_1$ for $\lambda x[\![app\,\langle[\![\cdot]\!]^\sigma\rangle\,x)]\!]$, and more generally $[\![\cdot]\!]_{n+1}$ for $\lambda x[\![app\,\langle[\![\cdot]\!]^\sigma_n\rangle\,x]\!]$. Thus more generally, CE^n is the theory stating that $[\![\cdot]\!]_k$ satisfies the axioms of CE for each $k \leq n$. E.g. C($\neg$) becomes:[74]

$\mathrm{C}_n(\neg)$ $[\![\langle\neg\rangle]\!]^n = \neg$

According to the theory CE^n, we have terms for $n+1$ meaning relations, $[\![\cdot]\!]_0, ...[\![\cdot]\!]_n$ that all satisfy the formal deciderata of a meaning relation, giving a partial articulation of semantic pluralism: the idea that there are multiple relations satisfying the expressing role.

Note that in section 5.4 we cast some doubt on the idea that in general, every relation 'means' bears a meaning relation to is also a meaning relation. We rejected this on the grounds that meaning relations can do better or worse at satisfying the expressing role, and that a meaning relation that barely satisfies the expressing role might not relate 'means' to a relation that satisfies the expressing role. Thus our informal motivation does not commit us to endorsing CE^n for arbitrarily high n. The intersection of all of these theories is consistent, as seen by a straightforward compactness argument, however, the specter of ω-inconsistency looms in the form of McGee's paradox (see McGee [29]), giving us independent reasons not to endorse CE^n for all n.[75]

A.4 Models

In the following we fix a set W, a designated world $w^* \in W$. We assign domains to each type as follows:

$$D^t = P(W)$$

$$D^e = \{c_1, c_2, c_3, ...\}$$

$$D^{\sigma\to\tau} = (D^\tau)^{(D^\sigma)}$$

Note that, to save ourselves a few lines, we have set things up so that the expression constants, $c_1, c_2, c_3, ...$ can intepret themselves.

In the following we define a class of models $\mathcal{M}$. Each model $M \in \mathcal{M}$ assigns each expression of $\mathcal{L}^\sigma$ to a denotation belonging to D^σ. In the case of the primitive expressions $\forall, \wedge, \neg, c_1, c_2, ...$ these denotations will be fixed across the models, so that a model is completely determined by the denotation from $D^{e\to\sigma}$ it assigns to $[\![\cdot]\!]^\sigma$ for each σ.

We have been treating $[\![\cdot]\!]^\sigma$ as a term of type $e \to \sigma$, taking an expression of type σ (a certain sort of individual) to a semantic value in type σ. What happens if we, for example,

[74] There are evident connections here with the truth theory FS (see Friedman and Sheard [17]), and the fragments FS^n defined in Halbach [22]. It should be no surprise that the consistency proof considered in the next section bears the marks of a revision sequence, echoing the consistency proofs for these theories.

[75] In this setting we may create a McGee like paradox by considering a fixed-point of the property $\lambda x\neg\forall_\sigma n(CN(n) \to n([\![\cdot]\!]^t)x)$. Here $\sigma = (\tau \to \tau) \to \tau \to \tau$, and $\tau = e \to t$, and CN is short for the property of being a Church numeral: $\lambda x\forall_{\sigma\to t}Z((Z0 \wedge \forall_\sigma n(Zn \to Z(suc\,n))) \to Zx)$, 0 is short for $\lambda f\lambda x\,x$, and suc for $\lambda n\lambda f\lambda x\,n(f)(fx)$.

attempted to apply $[\![\cdot]\!]^\sigma$ to a dog, or something else which isn't an expression of type σ? In that case we pick an arbitrary value, $d^\sigma \in D^\sigma$ to be the denotation, a choice which will again be assumed to be fixed across all models.

The following three functions, $all_\sigma \in D^{(\sigma\to t)\to t}$, $and \in D^{t\to t\to t}$ and $not \in D^{t\to t}$, are used below, and are defined by setting $all_\sigma(f) = \bigcap_{a\in D^\sigma} f(a)$, $and(p)(q)$ as $p \cap q$, and $neg(p)$ as $W \setminus p$.

Definition 5. *A* model, *M, consists of an element of $D^{e\to\sigma}$, $|[\![\cdot]\!]^\sigma|_M$, for each σ, subject to the constraint that $|[\![\cdot]\!]^\sigma|_M(\langle\alpha\rangle) = d^\sigma$ when α does not have type σ.*

A model uniquely extends to a type-indexed collection of mappings $|\cdot|_M^{\sigma\, g} : \mathcal{L}^\sigma \to D^\sigma$, where g is a variable assignment: a type-indexed collection of functions $g^\sigma : Var^\sigma \to D^\sigma$. In the following we omit type superscripts from the denotation function,

- $|x|_M^g = g(x)$
- $|\forall_\sigma|_M^g = all_\sigma$
- $|\wedge|_M^g = and$
- $|\neg|_M^g = not$
- $|c_k|_M^g = c_k$
- $|[\![\cdot]\!]^\sigma|_M^g = |[\![\cdot]\!]^\sigma|_M$
- $|\alpha\beta|_M^g = |\alpha|_M^g(|\beta|^g)$
- $|\lambda x\alpha|_M^g = a \mapsto |\alpha|_M^{g[a/x]}$

We denote the set of all models $\mathcal{M}$.

We say that M is a model of a sentence, ϕ, iff

$$w^* \in |\phi|_M^t$$

where w^* is the designated world. We say it is a model of a set of sentences iff it is a model of each member.

The jump operation is a function $J : \mathcal{M} \to \mathcal{M}$, that takes a model, M, and produces another model $J(M)$, in which the interpretation of $[\![\cdot]\!]^\sigma$ is determined by $|\cdot|_M^\sigma$.

Definition 6. *The jump operation is defined as follows*

- $|[\![\cdot]\!]^\sigma|_{J(M)}(\langle\alpha\rangle) = |\alpha|_M$ *if α has type σ*
- $|[\![\cdot]\!]^\sigma|_{J(M)}(\langle\alpha\rangle) = d^\sigma$ *otherwise.*

Theorem 1. *Let M be any model. Then*

$$J^n(M) \text{ is a model of } \mathsf{CE}^n$$

In particular, $J(M)$ is a model of CE for any model M, and thus CE is consistent.

B Models of Semantic Pluralism

In this section we consider formalisms that are a bit more friendly to the idea that there are multiple expressing relations. While we saw that $\mathcal{L}$ allowed us to define a finite sequence of semantic expressing relations $[\![\cdot]\!]_0, ..., [\![\cdot]\!]_n$ and consistently have axioms stating they each behave compositionally, our picture has been one in which there are a large, presumably uncountable, number of expressing relations. Our discussion here will be brief, indicating various natural formalisms without developing them fully.

There are several choice-points for the framework: I'll briefly discuss three. One is to theorize in terms of two type-indexed families of predicates:

- E^σ, of type $(e \to \sigma) \to t$, a predicate which informally states that its argument is an expressing relation for expressions of type σ.
- $C^{\sigma\tau}$ of type $(e \to \sigma) \to (e \to \tau) \to t$, a relation which informally states that its arguments are co-ordinated expressing relations.

In fact, E^σ can simply be treated as an abbreviation for $\lambda x C^{\sigma\sigma}xx$. The claim that co-ordinated expressing relations are compositional can be captured in this language. For example, the analogue of C(app) becomes:

$$\forall_{e\to(\sigma\to\tau)}X\forall_{e\to\sigma}Y\forall_{e\to\tau}Z(C^{\sigma\to\tau,\sigma,\tau}XYZ \to Z\langle\alpha\beta\rangle = (X\langle\alpha\rangle)(Y\langle\beta\rangle))$$

Here $C^{\sigma\tau\rho}XYZ$ is short for the conjunction $C^{\sigma\tau}XY \wedge C^{\tau\rho}YZ$.[76] We call the resulting language, with the expression constants and logical vocabulary described in section A, $\mathcal{L}^1$.

An alternative and slightly more readable formalism is to explicitly introduce names for the different semantic expressing relations. For example we can introduced indexed families of meaning functors, $\{[\![\cdot]\!]^\sigma_i \mid i \in I\}$, indexed by some set I. Call the resulting language, including the expression constants and logical vocabulary, $\mathcal{L}_2$. By indexing the expressing relations, we won't be able to quantify over expressing relations: a limitation that could be removed by treating the subscripts as arguments, allowing us to quantify into the index position. We can also mix and match these formalisms, by including both explicit names for the expressing relations, and the general expressing and co-ordinating predicates.

The final formalism employs a particular conception of the expressing relations, namely those relations similar to the semantic expressing relation. It involves bringing the notion of similarity employed informally in section 4.1 into the object language, and defining being an expressing relation of type σ as being similiar to $[\![\cdot]\!]^\sigma$. The language and models are little more complex, and are described in section B.2.

B.1 Compositional models

A model of $\mathcal{L}_1$ is defined as in section A, except that it is determined by an interpretation of $C^{\sigma\tau}$ for every σ and τ: in other words, it is a choice of element of $D^{(e\to\sigma)\to(e\to\tau)\to t}$ for each σ and τ. Call the set of all models $\mathcal{N}$. The jump operation $J : P(\mathcal{N}) \to \mathcal{N}$ is specified as follows:

- $|C^{\sigma\tau}|_{J(X)}(f)(g) = W$ if there is a model $N \in X$, such that

[76] Further book keeping axioms need to be included as well. For example, we should have a type indexed analogue of transitivity — $C^{\sigma\tau}XY \wedge C^{\tau\rho}YZ \to C^{\sigma\rho}XZ$ — as well as analogues of symmetry and reflexivity, and the claim that $C^{\sigma\tau}XY \wedge C^{\sigma\tau}XZ \to Y = Z$.

$f(\langle\alpha\rangle) = |\alpha|^\sigma_N$ whenever α has type σ, and $f(\langle\alpha\rangle) = d^\sigma$ otherwise.

$g(\langle\alpha\rangle) = |\alpha|^\tau_N$ α has type τ, and d^τ otherwise.

- $|C^{\sigma\tau}|_{J(X)}(f)(g) = \emptyset$ otherwise

We define a sequence of classes of models by setting $J^0(\mathcal{N}) = \mathcal{N}$ and $J^{n+1} = \{J(X) \mid X \subseteq J^n(\mathcal{N})\}$.

A model of $\mathcal{L}_2$ is defined as in section A, except we must select an interpretation for each member of the family $[\![\cdot]\!]^\sigma_i$. Call the class of all such models $\mathcal{N}$. The jump operation is now a function $J : \mathcal{N}^I \to \mathcal{N}$, taking an infinite sequence of models to a model. Given such a sequence $(N_i)_{i\in I}$ the interpretation of $[\![\cdot]\!]^\sigma_i$ in $J((N_i)_{i\in I})$ is determined by the interpretation function $|\cdot|_{N_j}$ of the jth model from the sequence.

Definition 7. *Let $(N_i)_{i\in I} \in \mathcal{N}^I$. The jump operation is defined as follows*

- $|[\![\cdot]\!]^\sigma_j|_{J((N_i)_{i\in I})}(\langle\alpha\rangle) = |\alpha|_{N_j}$ *if α has type σ*
- $|[\![\cdot]\!]^\sigma|_{J((N_i)_{i\in I})}(\langle\alpha\rangle) = d^\sigma$ *otherwise.*

We define $J^0(\mathcal{N})$ to be $\mathcal{N}$, and $J^{n+1}(\mathcal{N})$ to be $\{J((N_i)_{i\in I}) \mid N_i \in J^n(\mathcal{N})$ for every $i \in I\}$.

We let CE^n_i be the theory axiomatized by replacing $[\![\cdot]\!]^\sigma$ in CE^n with $[\![\cdot]\!]^\sigma_i$.

Theorem 2. *Every model in $J^n(\mathcal{N})$ models CE^n_i for every $i \in I$.*

B.2 Similarity

Here I resume some of the ideas discussed in section 5.5. We will investigate the thought that meaning relations are similar, and the idea that we can identify the property of being an expressing relation with the property of being similar to *means*. This requires us to expand our object language to be able to talk about perturbations. As before, we suppose that the set of worlds W comes with a metric $d : W \times W \to \mathbb{R}$, where $d(w, v)$ intuitively is some measure of how similar w and v are. Secondly we assume some suitably small radius $\alpha \in \mathbb{R}$ determining what it means for two worlds to be 'close'.

> A strong perturbation is a permutation $j : W \to W$ such that $d(w, j(w)) \leq \alpha$ for every $w \in W$.
>
> A weak perturbation is a (possibly non-surjective) injection $j : W \to W$ such that $d(w, j(w)) \leq \alpha$ for every $w \in W$.

Given a set $p \subseteq W$, write $j(p)$ for $\{j(w) \mid w \in W\}$. We can then say that two propositions, p and q, are close iff there is some perturbation, j, such that $j(p) = q$, or $j(q) = p$. Two notions of closeness are available depending on whether we use the weak or strong notion; we'll use the weaker notion in what follows. We may extend the action of a weak or strong permutation, j, to arbitrary types as follows. We make the simplifying assumption that the inhabitants of type e are all abstract objects, and are fixed by every perturbation: thus $j^e(a) = a$ for every $a \in D^e$ and permutation. We set $j^t = j$. Once we have determined what $j^\sigma : D^\sigma \to D^\sigma$ and $j^\tau : D^\tau \to D^\tau$, we fix some left-inverse of j^σ, i^σ, and define $j^{\sigma\to\tau} : D^{\sigma\to\tau} \to D^{\sigma\to\tau}$ as follows:

$$j^{\sigma\to\tau}(f) = j^\tau \circ f \circ i^\sigma$$

An inductive argument establishes that j^σ is injective for each type σ, which allows us to use assume the existence of a left-inverse. When we are dealing with strong perturbations it follows that j^σ is a permutation for every σ, and so this choice of left-inverse is uniquely determined and is simply j^σs inverse. In the case of weak perturbations there may be multiple ways to extend the action of j to other types.

In order to develop an object language theory of perturbations, we assume that perturbations are individuals. Indeed, our exposition is greatly simplified if we assume that every individual is a perturbation (thus some individuals play a double role of representing expressions of the language and representing perturbations).[77] For each type, σ, we have an action operation act^σ of type $e \to \sigma \to \sigma$, that lifts a perturbation, i, to an operation $act^\sigma i$ of type $\sigma \to \sigma$. When i is an individual, thought of as a perturbation, we will adopt the convention of writing i^σ instead of $act^\sigma i$. We may formalize our theory of perturbations by axioms: guaranteeing that perturbations are injective, fix individuals and logical connectives, and that they distribute over application

Pe $\forall_e ix\,(i^e x = x)$

P(inj) $\forall_e i \forall_\sigma xy(i^\sigma x = i^\sigma y \to x = y)$

P(app) $\forall_e i \forall_{\sigma\to\tau} f \forall_\sigma x\,(i^{\sigma\to\tau} f)(i^\sigma x) = i^\tau (fx)$

P($\wedge$) $\forall_e i\,(i^{t\to t\to t}\wedge = \wedge)$

P($\neg$) $\forall_e i\,(i^{t\to t}\neg = \neg)$

P($\forall_\sigma$) $\forall_e i\,(i^{(\sigma\to t)\to t}\forall_\sigma = \forall_\sigma)$

These principles are secured in a model by ensuring that for each $d \in D^e$, there is some perturbation j such that $|act^\sigma|(d) = j^\sigma$ for every σ. What about the principle, discussed in section 5.5, that expression relations are similar to their successors? We can capture this principle with the following principle

Closeness $\exists i(i^{e\to\sigma}[\![\cdot]\!]^\sigma_n = [\![\cdot]\!]^\sigma_{n+1}))$

When $n = 0$ this states that what 'means' means is similar to meaning. More generally, the successor of a meaning function is similar to it. Here sketch in outline a model that validates this principle for the case where $\sigma = t$, and $n = 0$. It is evident how to extend the argument for $n > 0$, but I do not know whether these sorts of models also satisfy Closeness when $\sigma \neq t$.

We assume the domains D^σ have been define as before, except that we make the additional assumption that the set of worlds, W, has the cardinality of the continuum, $2^{\aleph_0}$, and that $d^t = \bot$, the empty set of worlds. Let $\mathcal{L}_3$ be the result of adding the constants $aact^\sigma$ to $\mathcal{L}$. For technical reasons, we shall also consider a language $\mathcal{L}_3^+$ that has a constant c_d, for every element $d \in D^\sigma$. We assume that $\mathcal{L}_3$, not $\mathcal{L}_3^+$, is the domain of our representation function $\langle\cdot\rangle$ (and we may retain our assumption that D^e is countable). Let $\mathcal{P}$ be the set of models of $\mathcal{L}_3$ based on such a choice of domains, subject to the constraints in definition 5, plus the constraint on $|act^\sigma|$ mentioned above. Every model in $\mathcal{P}$ uniquely extends to a model of $\mathcal{L}_3^+$ by interpreting c_d with d. The jump operation, $J : \mathcal{P} \to \mathcal{P}$ is defined as in section A, with the additional constraint that the interpretations of act^σ are preserved:

[77] This assumption needs to be lifted if we want to theorize about languages with names for ordinary objects like mountains.

$|act^\sigma|_M = |act^\sigma|_{J(M)}$. Suppose, that we have partitioned the models in $\mathcal{P}$ by the relation $\sim$ of making the same sentences true, and chose some bijective association, g, between the worlds W and the equivalence classes $\mathcal{P}/\sim$. Since J preserves $\sim$, J determines a mapping from $\mathcal{P}/\sim \to \mathcal{P}/\sim$, and thus, via this association, a mapping $j : W \to W$. Since J is injective on $\mathcal{P}/\sim$, so is j on W. Now we treat j as a perturbation, and extend it so that it acts to arbitrary types, $j^\sigma : D^\sigma \to D^\sigma$ as described above.

Now we construct a model of CE, the Perturbation axioms and Closeness. First, we must define the interpretation of $|act^\sigma|$. We only subject it to the constraint mentioned above, further making sure that:

There is some $d \in D^e$ such that $|act^\sigma|(d) = j^\sigma$ for every σ

where j is the aforementioned perturbation determined by the jump operation.

Second, we must define $|[\![\cdot]\!]^\sigma|$ for each σ. In order to do this we first define a collection of functions, $F^\sigma : (\mathcal{L}_3^+)^\sigma \to D^\sigma$.

- $F^e(c_k) = F^e(c_{c_k}) = c_k$.[78]
- $F^t(\phi) = \{w \mid \phi \text{ is true in } g(w)\} \cap im(j)$
- $F^{\sigma\to\tau}(\alpha)(a) = F^\tau(\alpha c_a)$ for every $a \in D^\sigma$

We may show by induction that for $a \in D^\sigma$, $F^\sigma(c_a) = a$, establishing that F^σ is surjective.

$|[\![\cdot]\!]^\sigma|(\langle\alpha\rangle) = F^\sigma(\alpha)$ if α is a type σ term of $\mathcal{L}_3$.

$|[\![\cdot]\!]^\sigma|(\langle\alpha\rangle) = d^t$ otherwise.

Noting that F^σ is defined on expressions of $\mathcal{L}_3$ in virtue of being defined on expressions of $\mathcal{L}_3^+$.

We may now establish that $j|[\![\cdot]\!]^t_0| = |[\![\cdot]\!]^t_1|$. It suffices to show that $(j|[\![\cdot]\!]^t|)(\langle\alpha\rangle) = |[\![\cdot]\!]^t_1|(\langle\alpha\rangle)$ for each term α. Since j^e is the identity, so is the left-inverse of j^e, so that this amounts to showing that $j(|[\![\cdot]\!]^t|(a_k)) = |[\![\cdot]\!]^t_1|(a_k)$. In the case that α does not have type t, the LHS is $jd^t = j\bot = \bot$ and the RHS is $F([\![\langle\alpha\rangle]\!]^t) = \{w \mid [\![c_k]\!]^t \text{ is true in } gw\} \cap im(j) = \bot$ (since $[\![c_k]\!]^t$ is interpreted by $\bot$ in every model $g(w)$.) When α is a type t term, ϕ we must show that $j|[\![\langle\phi\rangle]\!]^t| = |[\![\langle[\![\langle\phi\rangle]\!]^t\rangle]\!]^t|$, i.e. that $\{jw \mid \phi \text{ true in } w\} \cap im(j) = \{w \mid [\![\langle\phi\rangle]\!]^t \text{ is true in } w\} \cap im(j)$, which follows from the fact that ϕ is true in $g(w)$ if and only if $[\![\phi]\!]^t$ is true in $g(jw)$ (because we chose j so that $J(g(w)) = g(jw)$).

References

[1] Miroslava Andjelkovic and Timothy Williamson. Truth, falsity, and borderline cases. *Philosophical Topics*, 28:211–43., 2000.

[2] Andrew Bacon. Can the classical logician avoid the revenge paradoxes? *Philosophical Review*, 124(3):299–352, 2015.

[3] Andrew Bacon. The broadest necessity. *Journal of Philosophical Logic*, 47(5):733–783, 2018.

[4] Andrew Bacon. *Vagueness and Thought.* Oxford University Press, 2018.

[78]Note that c_k and c_{c_k} are the only normal form terms of type e.

[5] Andrew Bacon. Substitution structures. *Journal of Philosophical Logic*, Forthcoming.

[6] Andrew Bacon, John Hawthorne, and Gabriel Uzquiano. Higher-order free logic and the Prior-Kaplan paradox. *Canadian Journal of Philosophy*, 46(4-5):493–541, 2016.

[7] Yehoshua Bar-Hillel. New light on the liar. *Analysis*, 18(1):1–6, 1957.

[8] Ross Brady. *Universal Logic*. CSLI Publications, 2006.

[9] Tyler Burge. Semantical paradox. *The Journal of Philosophy*, 76(4):169–198, 1979.

[10] Herman Cappelen and Ernest Lepore. *Insensitive semantics: A defense of semantic minimalism and speech act pluralism*. John Wiley & Sons, 2008.

[11] Donald Davidson. Radical interpretation. *Dialectica*, 27(3-4):313–328, 1973.

[12] Cian Dorr. To be F is to be G. *Philosophical Perspectives*, 30(1):39–134, 2016.

[13] Cian Dorr and John Hawthorne. Semantic plasticity and speech reports. *Philosophical Review*, 123:281–338, 2014.

[14] Matti Eklund. The liar paradox, expressibility, possible languages. In J. C. Beall, editor, *Revenge of the Liar: New Essays on the Paradox*. Oxford University Press, 2007.

[15] Hartry Field. Mental representation. *Erkenntnis*, 13(1):9–61, 1978.

[16] Hartry Field. *Saving truth from paradox*. Oxford University Press, USA, 2008.

[17] Harvey Friedman and Michael Sheard. An axiomatic approach to self-referential truth. *Annals of Pure and Applied Logic*, 33:1–21, 1987.

[18] Haim Gaifman. Pointers to truth. *The Journal of Philosophy*, 89(5):223–261, 1992.

[19] Michael Glanzberg. A contextual-hierarchical approach to truth and the liar paradox. *Journal of Philosophical Logic*, 33(1):27–88, 2004.

[20] Laurence Goldstein. A consistent way with paradox. *Philosophical Studies*, 144(3):377–389, 2009.

[21] H Paul Grice. Utterer's meaning and intention. *The philosophical review*, pages 147–177, 1969.

[22] Volker Halbach. A system of complete and consistent truth. *Notre Dame Journal of Formal Logic*, 35(3):311–327, 1994.

[23] John Hawthorne. Epistemicism and semantic plasticity. *Metaphysical essays*, 1(9):185–211, 2006.

[24] Allen Hazen. Contra buridanum. *Canadian journal of philosophy*, 17(4):875–880, 1987.

[25] J Roger Hindley and Jonathan P Seldin. *Lambda-calculus and combinators: an introduction*, volume 13. Cambridge University Press Cambridge, 2008.

[26] Harold T. Hodes. Why ramify? *Notre Dame Journal of Formal Logic*, 56(2):379–415, 2015.

[27] Pauline Jacobson. Towards a variable-free semantics. *Linguistics and Philosophy*, 22(2):117–184, 1999.

[28] David Lewis. *Convention: a philosophical study*. Harvard University Press Cambridge, 1969.

[29] Vann McGee. *Truth, vagueness, and paradox: An essay on the logic of truth*. Hackett Publishing Company Inc, 1990.

[30] Charles Parsons. The liar paradox. *Journal of Philosophical Logic*, 3(4):381–412, 1974.

[31] Graham Priest. *In Contradiction: A Study of the Transconsistent*. Oxford University Press, 2006.

[32] Arthur Prior. On a family of paradoxes. *Notre Dame Journal of Formal Logic*, 2(1):16–32, 1961.

[33] Arthur Prior. *Objects of thought*. Oxford, 1971.

[34] Sven Rosenkranz and Arash Sarkohi. Platitudes against paradox. *Erkenntnis*, 65(3):319–341, 2006.

[35] Keith Simmons. *Universality and the liar: An essay on truth and the diagonal argument*. Cambridge Univ Pr, 1993.

[36] Keith Simmons. *Universality and the Liar: An Essay on Truth and the Diagonal Argument*. Cambridge University Press, 1993.

[37] Nicholas J. J. Smith. Semantic regularity and the liar paradox. *The Monist*, 89(1):178–202, 2006.

[38] Scott Soames. *Beyond rigidity: The unfinished semantic agenda of naming and necessity*. Oxford University Press, 2002.

[39] Alfred Tarski. The concept of truth in formalized languages. *Logic, semantics, metamathematics*, pages 152–278, 1956.

[40] Alan Weir. Token relativism and the liar. *Analysis*, 60(266):156–170, 2000.

[41] Charles Henry Whiteley. Let epimenides lie! *Analysis*, 19(1):23–24, 1958.

[42] Bruno Whittle. Self-referential propositions. *Synthese*, 194(12):5023–5037, 2017.

[43] Timothy Williamson. Indefinite extensibility. *Grazer Philosophische Studien*, 55:1–24, 1998.

[44] Timothy Williamson. Everything. *Philosophical Perspectives*, 17(1):415–465, 2003.

[45] Timothy Williamson. *Modal Logic as Metaphysics*. Oxford University Press, 2013.

[46] Timothy Williamson. Semantic paradoxes and abductive methodology. In Brad Armour-Garb, editor, *The Relevance of the Liar*. Oxford University Press, Oxford, Forthcoming.

[47] Elia Zardini. Truth and what is said. *Philosophical Perspectives*, 22(1):545–574, 2008.