A drawback for substitutional arguments

Justina Diaz-Legaspe [a,b,*], Adam Sennet [c]

[a] *Sociedad Argentina de Filosofia Analitica (SADAF), Philosophy Department, University of Western Ontario, Canada*
[b] *The University of Western Ontario, Department of Philosophy, Room 3148, Stevenson Hall, 1151 Richmond Street North, London, ON N6A 5B8, Canada*
[c] *Philosophy Department, UC Davis, USA*

ARTICLE INFO

ABSTRACT

Competing theories on the semantics of group pejorative terms (also known as 'slurs') comprise both advocates and opponents to the Identity Thesis (IT), according to which these terms and their neutral counterparts do not differ in semantic value. In the opponents' camp, Christopher Hom has offered an argument based on substitution of slurs and neutral counterparts that both supports his semanticist approach and cast doubts on all IT-based approaches to slurs. We aim to point to a dilemma triggered by this argument based on evidence showing that substitution of some words (including but not restricted to slurs) for non-problematically co-referential pairs may fail to preserve truth-values in some linguistics contexts.

© 2021 Elsevier Ltd. All rights reserved.

1. Introduction

Theorizing about language requires one to immediately deal with a choice between (at least) two options: is the object of theory a body of shared knowledge concerning the relationship of words, phrases and their structure to interpretations or is it a study of the linguistic actions people undertake in order to interact with themselves and others? The difference is important but the two objects of study need not be in conflict: if anything, it is difficult to see in principle how people could engage in successful linguistic actions without shared linguistic knowledge. For example, speakers often engage in and interpret metaphorical utterances such as Groucho Marx's famous quip 'A hospital bed is a parked taxi with the meter running' as meaning that the medical industry is perversely incentivized to make money by keeping you in the hospital. This interpretation of Groucho's utterance doesn't seem to be rooted in the meanings of the individual terms and their structure (which, if taken literally, mean something absurd). However, it is correspondingly hard to see how we could ever construct a reasonable interpretation of the *utterance* rather than the *sentence* if we didn't know with Groucho that 'taxi' meant taxi, that 'parked' meant parked and that 'parked taxi' meant a taxi that was parked.

Luckily for those who like puzzles, there is no shortage of puzzling phenomena for which it is unclear which line of study is best suited for explaining them. A case that has generated much interest and which recently has enjoyed a surge of academic attention is the case of slur terms, especially group pejorative terms (GPTs). It is not disputed that slur terms have the potential to cause offence when uttered. The source of this potential, however, is notoriously difficult to pin down. Should we think of the offensive character of slurs is due to their representing their targets negatively and contributing offensive

---

* Corresponding author. The University of Western Ontario, Department of Philosophy, Room 3148, Stevenson Hall, 1151 Richmond Street North, London, ON N6A 5B8, Canada

E-mail addresses: jdiazlegaspe@gmail.com, jlegaspe@uwo.ca (J. Diaz-Legaspe), amsennet@ucdavis.edu (A. Sennet).

meanings to the phrases and sentences they are parts of? Or should we explain the offensive potentiality of slurs in terms of cultural taboos against using terms on account of their history and use as weapons to degrade their targets?

The first option may seem obvious: if an utterance causes offence, surely that is best explained by the meaning of the word or phrase uttered being offensive in meaning. But that conclusion is hasty: we know, for example, that some utterances are offensive even if they don't contain offensive parts. Claiming 'You also had people that were very fine people on both sides' can cause legitimate offence when uttered after a rally full of Nazis occurs, but not because the sentence has a particularly offensive part. Similarly, using a nickname someone disprefers shows disrespect and can cause offence without the nickname itself meaning something offensive. Other terms, like 'scumbag' we know to be pejorative because of what they mean, even if their use sometimes completely fails to cause any offence. So where do slurs fall? And what methodology could we use to investigate them?

The wide array of plausible answers in the literature can be narrowed down by throwing some light on the use of the theoretical terms 'semantic' and 'semantic content' in the debate. Most parties equate 'meaning' to information or truth-conditions, following the lead of the early works in analytic philosophy. For simplicity, we will use 'semantic content' to refer only to the contribution sub sentential components make to the truth-conditions of the sentence they are part of, as used in a given context of utterance. In this restricted perspective on meaning, some words simply lack a truth-conditional meaning; instead, they fulfill another communicational role with their use, like expressing emotional states or prescribing or commanding. For example, 'ouch' is clearly a word but not one that expresses a representation so much as signals that the utterer is in pain. Other words or phrases may have semantic content as well as the ability to communicate more than simply their contribution to a sentence's truth conditions. For example, he Spanish terms 'tu' and 'usted' both contribute the addressee to the truth conditions of sentence's they are parts of but they differ nonetheless in communicative role. With this narrow concept of semantics in play, the question on the meaning of GPTs can be stated more clearly: do slurring terms for groups have semantic values that involve negative properties characterizing their targets or is the relationship to these negative properties to be captured by means of other (non-semantic) tools?

This issue is clearly not the most important or pressing matter concerning slurs. Issues about how slurs are used as weapons of oppression, how they are reclaimed by targeted groups, and how they function within the ideologies of social systems have more social and ethical interest. Nonetheless, we think there is value in trying to get clear on what they mean. As noted above, a theory of language use is unlikely to be possible without a theory of shared meaning to partially explain how we coordinate on use. So we hope this paper contributes, even if indirectly, to investigations of those issues.

It is hard to say that any particular view has attained the status of orthodoxy on the topic, and intuitions are notoriously delicate. In particular it is often hard to separate intuitions about derogatoriness from intuitions about offensiveness. But any view that denies that GPTs have semantic values that crucially involve negative representations of their intended targets has to face a powerful line of argument advanced by Christopher Hom and Robert May.[1] Put roughly, swapping GPTs for purportedly synonymous non-GPTs (and no other terms) in sentences sometimes results in an apparent difference in truth value. Without good reasons to think appearances are misleading, we should trust that they are accurate.[2] Given plausible appropriate compositionality principles, a difference in truth value is determined by a difference in the meanings of the sentences' parts. Since the only difference in meaning can stem from the substitution of the GPT for the non-GPT, we should conclude that they must differ in meaning. Given the difference in meaning, it is not hard (but not entirely trivial) to argue that the difference in meaning consists at least partly in the lexical encoding of the aforementioned negative features in slurs' linguistic meaning -absent from the linguistic meaning of non GPTs. Thus, if you disagree with the idea that GPTs differ in semantic value from some more neutral ways of referring to the relevant group, you have your work cut out for you in deciding where things went wrong in the line of argument above.

This paper attempts to do some of that work. We will argue that the morals to be drawn from such substitutional arguments are less straightforward than appearances might suggest. In particular, we will try to show that these arguments over-generate in the case of terms that differ in features related to socially constrained language variations. We will conclude that substitutional arguments, at least as applied to the GPTs in question and as applied to socially marked lexical items in general, are either invalid or valid but not sound, and we will offer some explanations for where the arguments can lead us astray.

2. Substitution arguments

One platitude about language is that sentences (and therefore utterances) are complex entities, constituted at least in part by concatenations of words. Although certainly a platitude, this fact provides a launching point for a number of questions of interest for philosophers of language. Particularly, it does so because it allows us to create sentential frames by fixing some words in a sentence and moving out others. Consider for example another quip by Groucho Marx: ' morning I shot an elephant

---

in my pajamas. How he got in my pajamas I don't know.' The first sentence can be isolated and transformed into a sentential frame just by keeping fixed all words but the last:

(1) One morning I shot an elephant in my _____.

This simple idea has proven to be useful for many endeavours. For example, creating a frame that takes out all the words can show the syntactic structure of a sentence by indicating what kind of word can be inserted in each spot:

(2) <u>(Adverbial phrase) (subject: noun) (verb) (object: determiner + noun) (prepositional phrase)</u>

Sentential frames are also part of the first step taken in the direction of a certain theoretical path, ones that equates word meaning with the contribution words make to the overall meaning of the sentence. In a second step along this path, utterances of declarative sentences and worldly facts are bridged: the latter make the former true because the sentence uttered represent those facts: as it is drilled to philosophy students to the point of exasperation, utterances of the sentence 'Snow is white' are true (in this world at least) because of the fact that snow is indeed white (dogs and ice melts aside), and utterances of 'The cat is on the mat' are true only if the salient cat is on top on the salient mat, and false otherwise. These are often called 'truth conditions' Meanings of words are extricated from the truth conditions of sentences by paying attention to what they contribute to them. Thus, based on (1), we can pinpoint the meaning of the word 'pajamas' by inserting another word in the spot and taking a look into the truth conditions of the resulting sentence:

(3) One morning I shot an elephant in my pajamas.
(4) One morning I shot an elephant in my bedroom.

The difference in meaning between (3) and (4) is explained by the different contributions of the substituted words: while (3) is true if the speaker one morning shot an elephant in their *pajamas*, (4) is true if they did so in the *bedroom*. A correlation is thus established between words ('pajamas', 'bedroom') and worldly entities (pajamas, bedrooms). The ambiguity of (3), moreover, is due to the two different sentential frames it can be constructed from: one in which 'in my pajamas' is a an adjectival phrase, the other in which it is an adverbial phrase.

Sentential frames and substitution is the launching point for many philosophical arguments. Philosophers may remember substitutional arguments from Aristotle's infamous masked man argument[3] or from Frege's evening star/morning star cases.[4] The key idea is a powerful one: if your run-of-the-mill declarative sentence represents the world as being as such (at a time, place, world etc.), then another, similarly relativized, sentence purportedly representing the world in the same way but that (potentially) differs in truth value from the former must differ in meaning, because you cannot represent the same thing in the same way and differ in accuracy. Substitutional arguments should thus be useful in pinpointing differences in meaning between two expressions that seem synonymous at first blush.

Substitutional arguments follow a given structure: provided a sentence that acts as a frame, one of the constituents is substituted for another that purportedly shares the same semantic features. If the substitution results in a difference in meaning (as shown by a difference in intuitive meaning or truth value), there is evidence that the substituted expressions differ in meaning. The argument is grounded on the assumption of a variant of the Compositionality Principle:

Compositionality Principle: The semantic value of a sentence is determined by the semantic values of its parts and its sentential frame.[5]

It is plausible that sentential semantic values are compositionally determined. Thus, the substitution of two phrases or words with identical semantic values should not affect the semantic value of the sentence where the substitution occurs. If it does, it means one of these two things: either the substitutional word pair did not really have identical semantic values, or the sentence is sensitive to another semantic feature in which the substitutional pair differs.

Let us clarify through an example. Frege argues that names have meanings, not just reference. He does this using a substitution argument. Consider the co-referential pair 'Hesperus' and 'Phosphorus (both refer to the planet Venus). Because of this co-reference, truth-value preservation is to be expected through the substitution in (5) and (6):

(5) Hesperus is identical to Phosphorus.
(6) Hesperus is identical to Hesperus.

Both sentences are, as expected, true. However, something has changed from (5) to (6): (5) is informative and cognitively rich, while (6) is a tautology, not informative at all. Given that informativeness and cognitive richness depend on sentence

---

meaning (a highly controversial assumption!), Frege uses this result to posit the existence of a second dimension of meaning together with reference, *sense*. 'Hesperus' and 'Phosphorus' are co-referential yet differ in sense. Frege appeals to substitution for the second time to show a particular sort of sensitivity in sentences containing propositional attitude verbs:

(7) Adam thinks Hesperus is Hesperus.
(8) Adam thinks Hesperus is Phosphorus.

Again, although both of the sentences 'Hesperus is a planet' and 'Phosphorus is a planet' considered in isolation have identical truth-values, it could be the case that Adam associated two different senses with 'Hesperus' and 'Phosphorus' respectively, making (7) and (8) potentially different in truth value. From this divergence in truth value Frege concludes that verbs like 'think' cause a shift in reference for the terms in their scope. All this is to say that substitution arguments make a powerful cases for difference or sameness of meaning.

Of course Frege's diagnosis is only one amongst many, and not a popular one these days.[6],[7] This multiplicity of diagnoses shows something important: substitution arguments can show a need for an explanation. They can provide evidence for a difference in sentence meaning due to the substituted word or phrase. They generally do not, however, determine a unique explanation for the phenomenon they exhibit.

Given that substitution arguments can give evidence for differences in meaning, we can presumably use them to help make progress on the puzzle of GPTs. Christopher Hom pursued exactly this line of argument in his influential (2008). We turn to this next.

3. Hom's case

Our interest in substitutional arguments concerns their use in the literature about GPTs. In what follows, we will focus on the way Hom (2008) uses substitutional arguments to show a difference in semantic value between GPTs and purportedly synonymous non-slurring terms.

GPTs are pejorative words demeaning groups of people in virtue of their race, ethnicity, religion, nationality, gender, and sexual orientation, among others. Current debates about the meaning of GPTs have been particularly heated. All parties agree that GPTs seem to be used to talk about a particular group, and that predicative uses are taken to have a particular group in their extension. But they are unlike other words for the same group in the fact that they carry 'derogatory force': they derogate groups with such virulence that they are extremely offensive and damaging. Moreover, derogatory force seems to be a linguistically encoded feature of these terms. It is notoriously easy to show this by comparing word pairs like 'kike' and 'Jewish person,' 'spic' and 'Hispanic,' the N-word and 'black people.' Philosophers of language have offered several ways to accommodate derogatory force within different accounts of meaning for these words.

We have no shortage of tools at our disposal to appeal to for explaining the source and nature of derogatory force. A natural thought about derogatory force is that words like 'kike' are derogatory terms for Jewish people, in part, because they refer to Jewish people. These philosophers endorse the following thesis:

Identity Thesis (IT): GPTs and their neutral counterparts do not differ in semantic value.

Endorsing IT does not commit one to thinking that GPTs aren't derogatory or that they aren't important. It just means that GPTs *are* derogatory in virtue of something other than how they *represent* their targets.

Thus, some philosophers claim that derogation is due (entirely) to a non-optional conventional implicature,[8] or even a special kind of conversational implicature.[9] On this line of thought, there are two terms to refer to Jewish people and there is a sort of agreement amongst users of English that one of them is neutral and the other is only used to refer to Jewish people while betraying a dislike of them. Others argue that GPT meaning has an extra expressive layer allowing users to express negative attitudes towards groups referred to by the GPTs.[10]

By contrast, some theories take the distinction between GPTs and neutral counterparts to be grounded in their semantic values. From this perspective, GPTs and neutral counterparts differ in reference (extensionally) and the property picked out (intensionally): GPTs don't refer to any group of people at all since they literally mean 'person in group G who by virtue of being in G is bad in some way' and no-one is bad just by being a member of a group.[11],[12],[13]. The semantic value is typically

---

[6] For example, Crimmins and Perry (1989).

[7] Most notoriously, Salmon (1986 and 1989) and Soames (1987 and 1988).

[8] Among others, Sennet and Copp (2015), Whiting (2013), and Williamson (2009, 2010).

[9] See Nunberg (2018).

[10] Jeshion (2013) and Camp (2013).

[11] Hom (2008, 2010, 2012) and Hom and May (2013, 2018).

[12] Croom (2011, 2013).

[13] As one of the anonymous referees has pointed out, this depends on what is considered a group in this context. We won't test the reader's patience with our notion of 'group', but here we refer to groups of people singled out by nationality, race, religion, immigration status, etc., that is, the groups of people referred to by slurs. We thank a referee for the observation.

taken to be determined by stereotypical properties generically associated with the target group, while neutral counterparts have their expected meanings. A sophisticated externalist version of the null-reference version of these theories, which we will call 'non-IT approaches,' is proposed and defended by Hom (2008) and Hom and May (2013).

There might be nuances, but the rivalry between non-IT and IT-based approaches to GPT meaning can be framed thus: either GPTs and neutral counterparts differ in semantic value or they do not. This may not be the most important philosophical issue concerning GPTs, but it is one that captured the attention of those working on philosophical semantics and meta-semantics. What is interesting for us is that substitutional arguments have been claimed to settle the issue, which we will turn to next.

Hom (2008) puts forth interesting substitutional arguments to show that the IT is false.[14] Recall that endorsing (IT) commits one to seeing GPTs and their neutral counterparts as equivalent in semantic value. Endorsing (IT) and compositionality commits one to thinking that one shouldn't expect a difference in sentence meaning between sentences that differ only in containing the GPT in one case and the neutral counterpart in another.

Hom tests this result by consideration of sentences such as (9–14):

(9) John is anti-Semitic because he thinks that Jews are kikes.
(10) John is anti-Semitic because he thinks that Jews are Jews.
(11) It's anti-Semitic to think that a kike lives next door.
(12) It's anti-Semitic to think that a Jewish person lives next door.
(13) Institutions that treat Jews as kikes are anti-Semitic.
(14) Institutions that treat Jews as Jews are anti-Semitic.

(9) seems clearly true, but (10) seems clearly false. (11) seems true, and then (12) seems false. While it is not altogether clear what it means 'to treat Jews as Jews' (a point we return to below), (13) seems true, and (14) seems false. Hom concludes that 'Jews' and 'kikes' do not share their meaning and *mutatis mutandis* for other such pairs of GPTs and non-slurring counterparts. This conclusion supports his theory that non-GPTs like 'Jew' have Jewish people as their extension while 'kike' (and all GPTs) fail to refer or to have a non-empty extension.

One may be tempted to contest these truth value assignments: clearly there are ways of uttering (10), (12) and (14) that will exhibit a speakers contempt for Jewish people (and correspondingly for other sentences with GPTs). For example, we suspect you can imagine someone uttering (10) with contemptuous intonation on the second occurrence of 'Jews.' Contrastive stress or derisive pronunciation of 'Jew' can easily impact the message that is conveyed by each of the utterances above. This is an interesting point and raises a methodological question about our capacity to evaluate sentences rather than imagined utterances of those sentences – as we've been pointing out, utterances can convey more information than the literal semantic content of the sentence uttered. However it is pretty clearly the derisive intonation rather than the meanings of the words being stressed that is at issue. Derisive intonation can be used on just about any word to exhibit the speaker's dislike and contempt for the group. As such, we will leave then these non-semantic intrusions out of our considerations, however interesting they may be, and trust the reader to focus on the meanings of the sentences in question, not on creative ways in which they can be uttered.[15]

Hom's general line of thought is, roughly, that GPTs have their meaning determined partly -and externally- by institutional ideologies regarding the relevant groups, while non-slurring terms have their semantic content determined without reference to such ideologies. It is an interesting question why the GPTs are sticky enough to pick up residues of ideology while terms referring to the same targeted group do not, given that the same ideologies exist relative to the same groups. Hom and May (2013) explain this as an application of an operator ('PEJ') that applies to expressions, making them slurring expressions, and returns the null set as reference. Fortunately, the details need not concern us here. The main question in this paper concerns the viability of IT-based views in general, not the correctness of any given non-IT view.

The substitutional argument thus plays a key role in the dispute. If, as IT claims, GPTs and their neutral counterparts have the same semantic value, substitution should not yield a difference in truth value. So we have an apparent dilemma: either (9–10), (11–12) and (13–14) do not differ in truth value despite appearances (in which case our intuitions are unreliable) or they do, which is inconsistent with approaches that validate IT (assuming compositionality). In this context, simply denying the appearances looks like desperate pleading and as such, IT-based theories seem to be in trouble.

---

[14] Consider for example this excerpt from Hom and May's (2013): "*Because each sentence of the pair differs only in the inter-change of the pejorative and its non-pejorative counterpart, their resulting difference in truth-value is evidence that pejoratives make truth-conditional contributions to sentences in which they occur. Critically, those truth-conditional contributions must be derogatory for the correct result in the overall truth-value of the sentence. This indicates that postulating expressive content is superfluous.*"Expressive content is their word for aspects of meaning that aren't part of the truth conditions but are meant to express, rather than represent, the targets of the slurs in a negative light.

[15] We thank the anonymous reader for this observation. The gestured at problem is a general concern: using the tools language offers is a creative affair and we are very good at uttering sentences in order to convey very different things that they literally mean. As such, it is difficult to separate intuitions about a sentences (at a context) from intuitions about what someone might try to convey by uttering the sentence. Difficult but we trust not impossible. As we argued briefly above, the ability to convey more than a sentence literally means by uttering it depends on it literally meaning something and, moreover, that competent speakers' access to the meaning is at least possible if sometimes difficult.

But the philosophical point at issue goes beyond a theory of GPTs, as it also impacts on methodological issues in philosophical semantics. The intuition IT-based theories began with was that terms directed at a group refer to the group and that the nastiness of GPTs was to be explained by non-semantic means. The idea that we should take into account people's referential intentions and intuitions about reference as primary, defeasible evidence in determining the semantic value of a term has been a powerful guide in late 20th century approaches to meta-semantics. One might think of this as a 'reference first' approach to semantics. Hom's arguments, and the subsequent developments by Hom and May, suggest an error theory about GPTs reference, suggesting that in the relevant cases we look towards intuitions about reference determination of a term in order to see if the relevant term refers. One might think of this as a sort of 'sense first' approach.

Hom's use of the substitutional argument threatens the truth of any view committed to the IT, and if correct, dispositive; we can shelve an entire, highly influential, approach to the semantics of slurring. Moreover, it would leave us with an interesting puzzle: in what sense is a term like 'kike' derogatory towards Jewish people if it literally refers to no one whatsoever? The argument, in our view, deserves more attention.

So is he correct?

4. Failure of substitutional arguments

We think, and will argue below, that the answer is 'no.' This rejection may come as a surprise, for his cases are moving and powerful.[16] Our strategy will be to argue that the linguistic contexts in which the relevant substitutions take place fail to preserve apparent truth value even when words that are far less controversial and politically loaded than GPTs are at question. We will then argue that this forces a trilemma: either (i) these differences really are semantic differences, or (ii) the linguistic contexts Hom uses in his argument are sensitive to these words, and hence they are not good tests for substitution, or (iii) our intuitions regarding substitutions are unreliable, taking the dialectical sting out of Hom's line of attack.

We will begin by expanding the considerations to a greater range of substitution cases. The cases we will consider involve apparently synonymous pairs that differ in non semantic features. Some pairs could be constituted by dialectal or sociolectal alternatives ('restroom'/'toilet'), some pairs include a neologism or an archaism ('pickpocket'/'cutpurse'), and some include a difference in formality ('shit'/'poo'). We will take all of differences to reflect a difference in register. The notion of register is complex but the basic idea is that many synonymous terms differ in how appropriate it is to use them in different contexts. 'Boobs' and 'breasts' may mean the same thing but doctors examining patients generally refrain from using the former and risk sanction if they don't. We take it that *none* of these differences are differences in semantic value.

We are interested in whether or not substitutions of apparently synonymous word that differ in non semantic features can nonetheless result in a difference in intuitive truth value assignments. Particularly, we are worried about whether or not these differences occur in the same sorts of linguistic contexts Hom uses to test for semantic differences between GPTs and non-GPT counterparts. Our aim is then twofold: first, to show that his arguments over-generate and that thus, the data he appeals to should be explained using some other means. Second, given that GPTs are marked as different in register than related non-perjorative terms, if terms that differ in register *generally* cause problems in substitution arguments, it is no surprise that they do so in Hom's as well. Our general contention is that register, as well as the rest of sociolinguistic traits mentioned above, can play a role in fixing other potentially context-sensitive features of interpretation and that this is what is resulting in the differences in (9)–(14) above.

Consider:

(15) John is old fashioned because he hopes they catch the cutpurse that stole his wallet.
(16) John is old fashioned because he hopes they catch the pickpocket that stole his wallet.
(17) It's immature to be excited about taking a dump.
(18) It's immature to be excited about excreting.
(19) Doctors who tell people that their loved ones have died are insensitive.
(20) Doctors who tell people that their loved ones kicked the bucket are insensitive.

There is an intuitive change in truth value, one issuing from the presence, in the same sentence, of those key words ('old fashioned,' 'immature,' and 'tell') and words with different registers like 'shit' (vulgar register), 'cutpurse' (archaism), 'take a dump' (vulgar) and 'kick the bucket' (slang). (15) looks like a good description of John, if we focus on the way his hopes are articulated in the sentence, while (16) is not; even if we agree on what the rules of etiquette are when it comes to relieve ourselves when attending a dinner party, we find more people willing to agree with (17) than with (18); finally, we may find doctors insensitive in (20) because of the way they deliver bad news, even if we consider that task as part of their métier.[17]

---

[16] Moreover, most (we hope all) in the philosophical debate who defend IT-based approaches are a bit embarrassed to defend a view on which utterances of sentences like 'Jews are kikes' are true. We are, and we believe the thesis!.

[17] For each of these pairs of cases, one can for a purely referential reading of 'old fashioned', 'immature' etc. If that is how the above sentence pairs are read, the pairs share their truth-values. But we are confident that there is a prominent reading of each pairing in which they differ in truth value, which we discuss above. Our point is that the non-semantic features of these terms (such as a term's being old fashioned) can impact our intuitive assessment of their meaning and so long as there is a legitimate reading on which the pairs differ in intuitive truth value, we have what we need to make our point.

We are fine with the fact that only some particular readings of these sentences lead one to sense a truth value divergence; what we care about is whether or not there is a pair of readings on which these ways of thinking thoughts can diverge. And we think that clearly they can diverge even if there are available readings on which they do not. A fancier way to emphasize the discordant reading is by merging both sentences into one:

(21) John isn't old fashioned because he hopes they catch the pickpocket who stole his wallet; he's old fashioned because he hopes they catch the cutpurse who stole his wallet.

One may complain that these cases are likely to be cases of metalinguistic negation, in which the use of a word is implicitly challenged rather than as a denial of its content.[18] We agree, and we add that perhaps something similar is going on in (15)–(19) even though they do not include an obvious objection to the application of the key words. But if this approach is right, surely the same explanation can apply to Hom's cases. If 'anti-Semitic' as applied in (8) and (10) can target the means by which the thought is articulated, rather than the content of the thought, it is not surprising that Hom gets the results he is looking for. This result should be expected once you throw words with funny non-semantic features to the mix.

Cases like (13) and (14) are slightly different, as they rely on the linguistic context in the scope of 'treats like' and 'treats as.' Phrases like 'treats like' impose a comparison between different types of treatments. Comparisons, famously, are context-sensitive and/or underspecified, and thus it is not surprising that pragmatic or social aspects of the choice of terms can resolve the context-sensitivity and/or under-specification differently. In turn, this could affect the truth value of the sentence stating the comparison in the context of utterance. Comparing a person to a pig, for example, can underscore lack of table manners in some contexts, and in others their famously rotund bodies. A similar phenomenon seems to happen with phrases such as 'treat like.' For example:

(22) Institutions that treat water like H20 are scientific.
(23) Institutions that treat water like water are scientific.
(24) Don't invite people to your house if they treat your toilet like a toilet.
(25) Don't invite people to your house if they treat your toilet like a shitter.
(26) They treat his work like a collection of childlike ideas
(27) They treat his work like a collection of youthful ideas.

As happened with our (15–20) cases, there is clearly a reading of each of these pairs on which they may differ in truth value. Of course, one may claim that the above examples are proof that the substitutional pairs involved are non-co-referential. But we would obtain the same results even with a stipulated co-referential pair. Consider the term 'philosopherer': we stipulate that the term is co-referential with 'philosopher' but differs in the fact that its use displays an insulting attitude towards philosophers. On this basis, we take it that you will understand the following instructions differently:

(28) Treat him like a philosopher.
(29) Treat him like a philosopherer.

It thus seems like a poor diagnosis to foist the responsibility for this shift on a change in semantic value between 'H20' and 'water', 'toilet' and 'shitter' and 'childlike/childish' and 'youthful' respectively. A much more mundane and plausible explanation focuses on the sensitivity of 'treat like' to the way content is articulated, rather than to content itself. More generally, contexts that prompt comparisons are sensitive to non-semantic aspects of word choices. Again, this threatens the probative value of Hom's cases (13–14), due to the presence of 'treat like' governing over 'kike' and 'Jew', which differ in non semantic features (at least). Even if 'treat like' contexts are not sensitive to non-semantic differences, and these sentences are perfectly equivalent in every way, then we have reasons to doubt our intuitions about the truth values of the sentences involving the substitutions in question. But this line of thought applies to Hom's case just as much as to ours.

5. Sensitivity

Thus far, we have pointed to cases where non semantic features of words could interfere with the preservation of truth values in some applications of substitutional arguments. We took as given that both constituents of the substitutional pairs in each case have the same semantic value. Obviously this could be challenged but we think that it is a pretty desperate ploy. We will dive deeper into the differing features in each substitutional pair. Consider again the word pairs used in the cases above: 'cutpurse' and 'pickpocket,' 'take a dump' and 'excrete feces,' and 'die' and 'kick the bucket.' Assuming their synonymy, they still have clearly distinct socio-linguistic features: they all coincide in them being anchored in social, non semantic factors. The case of 'cutpurse' involves a term that is clearly an archaism, not in current use. Registered words also belong to this family of socially-marked words: vulgar and clinical ('take a dump'/'excrete') seem synonymous but appropriate to different

---

[18] See Horn (1985) for discussion of metalinguistic negation.

conversational settings. Last, sociolectal variations lead to differences like those between the slang and neutral variants 'kick the bucket' and 'die,' with differing frequency of use among speakers in different social classes or cultural segments. Notice that the social dimension of these words is considered *semantically negligible*; they interact with the participant's understanding of the type of conversation they are having, but they do not affect the semantic value of the words.

Despite being semantically negligible, substitution of words by socially marked co-referential alternatives may have a semantic impact on sentences, failing to preserve truth-value. We saw above that this can happen when sentence-frames contain terms that are sensitive to the modes of expressing content ('old-fashioned', 'inconsiderate') or involve a comparison ('treat like'). This could even happen in plain sentences (15–16), when the sentence is read in a certain way that focuses on the *manner* in which the information is presented and not solely on the information presented. We would say, then, that the presence of some phrases or syntactical constructions makes sentences *sensitive* to the social burden of the word.

The reader might be curious about the reason why some sentence frames are sensitive to socially-laden words. We do not pretend to have a full theory that explains these interesting facts, but we will try to point in the direction of some plausible explanations.

One of them points out to the potential hyper-intensionality of these sentential frames. In this picture, social-sensitivity is yet another flavor of hyper-intensional contexts, with sentences being more about names or ways of naming things than about their apparent content. This may work for some of our examples: for example, the way doctors articulate the news when they inform relatives of the demise of a patient certainly expose them to being judged as insensitive. However, the explanation does not fit other examples as nicely: for example, being inconsiderate seems to depend on what we *do*, not much on the words used to describe what we do, and even so we sense a truth value difference between (19) and (20).

Certainly, an advocate of the hyper-intensional approach can overcome these bumps on the road one way or another. However, at the end of the day appealing to hyper-intensionality seems to be more of a re-labeling of the phenomenon than a real explanation.

Another alternative builds upon the metalinguistic explanation, based on the concept of mixed quotation. The foundation for this explanation lies in noticing that quotes are used for a multiplicity of reasons other than citing.[19] In this case, quotes could be bringing to the fore the social marks of these words. Thus, a proper representation of the resulting sentences would contain the quotes that we do not see but should hear. Again, this seems plausible for some of our cases:

(15*) John is old fashioned because he hopes they catch the "cutpurse" that stole his wallet.
(20*) Doctors who tell people that their loved ones "kicked the bucket" are insensitive.

Quotation is sometimes responsible for creating hyper-intensional contexts. The quotes are used here to indicate the social burden of the word: in the first case, that it is an outdated expression, and in the second, that it is a slang, somewhat vulgar phrase. Underscoring these phrases works well with the perceived intention of the speaker: to point out at how 'insensitive' some doctors are, and how 'old fashioned' is John. However, this explanation does not seem to help much with the cases in (22)–(29). They at least do not seem to be meta-linguistic at first blush.

There is, nevertheless, a line of thought that we favour in explaining the data. Phrases like 'thinks that,' 'treat that,' 'being immature,' 'being crude' and others, seem to be sensitive to *how* the subject is thinking of the relevant object, not only to the fact that they are thinking about the object itself. If they are so, the reason why substitution of socially marked phrases impacts on intuited truth values in the cases above is because these words indicate a way under which an object is being thought of. To think about, say, feces from the perspective of a physician role differs from thinking about it from the typical perspective of a child, even though both thoughts are about the same type of object and could even share the same content. Likewise, the presence of expressions like 'crude,' 'old fashioned,' 'insensitive,' and 'immature' also seem to call out for paying attention to the manner in which the thought is being entertained. Thus, the change in social mark in each case calls attention to an aspect of the vehicle the thinker in question is using to represent the content.

Readers may be curious as to the story behind the phenomenon. Ideally, it could involve resorting to an association between words and non-semantic mental objects called 'Ways of Thinking' (or WOTs). Copp and Sennet (2017) claim that WOTs are conventionally connected to pejorative terms. This account seems to accommodate the differences between GPTs and non-GPTs counterparts that are relevant for Hom's argument. According to this approach, the WOTs associated with GPTs would involve thinking about the target group and its members as contemptible, maybe in a stereotypified way, features not present in the way the same groups and individuals are thought of under the WOTs associated to non-GPTs counterparts. It is not difficult to extend this idea to cover other kinds of social marks too. The idea behind a WOT is that it characterizes the differences between how, say, a chemist and a child think of water. They can co-ordinate their thoughts because they refer to water, but the differences in the overall mental orientation towards water between the two is taken to be a WOT, including the general level of knowledge about the substance, attitudes towards it, stereotypical instances of it, etc. When reporting thoughts for practical purposes, it is not surprising that the WOTs matter, as the chemist and the child above who both believe that there is water in the river will typically be disposed to act very differently if asked if there is H2O in the water. Taking this idea one step further, socially marked words and non-marked co-referential terms would differ in the associated WOTs,

---

[19] Predelli (2003).

triggering the intuited failures of truth value preservation under co-referential substitution in the socially sensitive contexts mentioned in this paper.

A dovetailing but unexplored source of evidence may come from studies in psycholinguistics, which provide experimental evidence that people hold strong associations between words in clusters. Subjects of this kind of experiments are exposed to a particular word ('apple' and are prompted to offer another word in quick association (typically, subjects elicit words like 'red,' 'sweet,' 'fruit,' etc.). According to the results, words seem to cluster in groups tightly connected to each other in what the authors call 'stereotypes'.[20] These stereotypes, they claim, shape interpretative processes and can affect inferences one makes that aren't grounded in a word's meanings but in developed associations. For example, it is plausible that contemporary speakers tend to associate 'secretary' with 'woman.' This would explain why use of the term 'secretary' tends to lead to an inference that the secretary is female, causing a need for clarification to the contrary if the secretary in question is male. If this is right, then synonymous terms could be associated with different clusters triggering different inferences. Consider for example 'bunnies are cute' and 'Oryctolagus cuniculi are cute.' We suspect that the association of cuteness triggered by 'bunnies' is more immediate than the latter. Of course, being cute is part of the *meaning* of neither – the sentence 'that bunny is not cute' isn't automatically false because of its meaning. Perhaps similarly we can use this phenomenon to help explain our judgements in the substitution cases as a confusion of association based inferences for part of the meaning of terms. We'll leave this to further research.

6. Consequences for Hom's substitutional argument

If we are right and some of Hom's key cases are sensitive to substitution of socially marked terms, we have reason to be suspicious of their use in diagnosing the semantic contribution of GPTs. After all, all contenders in the debate over GPTs agree that GPT and neutral counterparts do not belong to the same register category.[21] As mentioned earlier, the concept of register comes from sociolinguistics,[22] and it names language variations that depend on the type of social setting or institutionalized situation in which language is used. Perhaps, then, the non semantic difference between 'kike' and 'Jew' is the same that exists between 'poo,' 'excrement' and 'shit.' GPTs are socially marked, in the same way as other lexical items are socially marked by speaker's provenance (dialect) or social status (sociolect), to mention a few.

Our argument in a nutshell is that in virtue of their social mark, substitution of registered words may result in a truth value difference in some sensitive sentential frames. To add a few more cases:

(30) Adam is immature because he thinks that excrement is poo.
(31) Adam is immature because he thinks that excrement is excrement.
(32) It's crude to think that there is shit in the toilet.
(33) It's crude to think that there is excrement in the toilet.

Presumably the differences in truth value don't reflect a lack of common semantic value in 'poo' 'shit' and 'excrement.' This would amount to a fairly drastic error theory about the terms. Presumably too, the occurrence of 'poo' and the common knowledge that it is a childlike way of referring to excrement interacts with 'immature' and similarly for 'shit' and 'crude,' the result of such interaction being the intuited difference in truth values.

A similar problem arises for Hom's 'treat like' cases. We hypothesize that 'treat like' is generally semantically under-specified (or interestingly context sensitive), which reflects in the fact that there is no obvious out-of-context sense in which we can interpret 'the institution treats Jewish people like Jewish people.' It is much easier to fix a range of possibilities in context: for example, if the institution where the sentence is uttered is a restaurant, plausibly 'treat like Jewish people' relates to possible dietary restrictions, while if it involves an institution holding a lecture about the rise of anti-semitism in the US it may involve sensitivities to background level of knowledge or experience. As mentioned above, there may be a connection between this feature and the way we think about objects under different semantic variants. Thinking about (Jewish) in-dividuals as Jewish differs from thinking about them as 'kikes,' and both anchor different ways of treating them in different contexts and respects. Obviously, this is hard to specify out of context but, and here is the key point, it wouldn't be surprising if the choice of terms to use in specifying the person/group/event/property in the second argument place of 'treat x like y' helped alter the intended in-context range of treatments. It is thus not surprising that features other than reference of the chosen term can help make a real difference to the in-context interpretation of the sentence. In general, we use all and any clues available to help specify a meaning when specific meanings are elusive, and it seems downright rational to use rather than discard all kinds of linguistic information in the process, including social marks and non semantic traits.

Again, this reinforces the idea that 'kike' and 'Jew' may well make different contributions to sentence interpretation *even if* they share a semantic value, when other constituents of the sentence depend on context for their interpretation and word choice helps fix the relevant context. Differences in register trigger sensitivity to social marking in the same way as do words marked by dialect, sociolect, chronicity, etc. Substitution of registered words, hence, are also subjected to the same objections

---

potentially raised against our cases above. Thus, Hom's (5–10) cases could be explained by appealing to the use of quotes, as cases of metalinguistic remarks. They can also be explained as instances of hyper-intensional contexts. In any case, these uses of the substitutional argument on words bearing a social mark do trigger at least one reading in which truth value is not preserved.

Cases involving registered substitution in contexts sensitive to social traits like the ones above pose a dilemma for Hom's substitutional argument. On the first horn, a trusting attitude on these cases would lead people to accept that some words create these special sensitive contexts, that is, sentential frames in which socially marked substitution does not preserve truth value. If these cases exist, Hom's (9–14) substitution is definitely one of them. As such, it ceases to be a convincing argument against IT: after all, us trusting philosophers would claim that he has chosen a sentential frame where substitution of socially marked words won't preserve truth values to prove that registered substitution – a type of social marking – won't preserve truth values. More precisely, the failure in truth value preservation in these cases is not due to an extensional divergence between both terms, but to the social sensitivity of the expression 'treat sth as sth,' or to the sensitivity of key words in some of these sentences to social, non semantic features of words. IT-based approaches, like pragmatic and expressivistic approaches to GPTs, are off the hook, for intensional divergence is not responsible for the failure in truth value preservation.

Of course, not every philosopher is so trustful. On the second horn of our dilemma, a more skeptical attitude on the above cases would seek for an alternative explanation of what is going on in them. As hinted in the last section, there could be a myriad of different explanations as to why some words or word combinations create socially sensitive contexts, more likely involving some sort of pragmatic or presuppositional explanatory device. Skeptical philosophers in this second horn of the dilemma would claim that, of course, extensionally divergent words fail to preserve truth values when substituted in an appropriate sentence frame. However, in *these* cases, some pragmatic phenomenon is blocking truth value preservation: there is no such thing as "social sensitivity." But even this skeptical perspective puts Hom in a bad place: the example he provides for his substitutional arguments exemplify whichever pragmatic obstacle the skeptical philosopher is pointing at, and then, it is not a reliable example to prove IT-based approaches to GPTs are wrong.

Here is our reasoning and a last concern:

P1. Word pairs like 'poo' and 'feces' (and the other cases discussed above) don't differ in reference.
P2. 'Adam is immature if he thinks there's poo on the floor' and 'Adam is immature if he thinks there's feces on the floor' apparently do differ in truth-value (on the relevant reading).
P3. If they do differ in truth value, then replacement of co-referencial terms can fail to preserve truth value in socially sensitive environments.
P4. If they don't differ in truth value, then our intuitions about substitution of co-referentials that differ in social marking are unreliable (since they seem to differ in truth value under substitution)
P5. They either do or don't differ
C. The substitution argument doesn't preserve truth-values, or our intuitions are unreliable.
C1. If (C) then the substitution arguments for differences in reference for GPT and their counterparts is either invalid or our dialectically impotent against IT-based theorists.

We expect challenges to (P1) and (P2). It's hard to *prove* (P1) but we find it plausible, and the reasoning against that plausibility ought not, we think, stem from substitutional arguments. We hope you find (P2) plausible for at least some of the cases presented above. In turn, (P3) raises concerns that we need to deal with here. You might worry that if substitution of co-referential phrases can change the truth value of a sentence then disaster looms: the Compositionality Principle may be put to question. If you replace a term with a term that make identical contributions to the truth value of the sentence, surely compositionality dictates that the resulting sentence meaning is identical. Doesn't endorsing (P3) commit one to denying compositionality, a property which enjoys such a high degree of theoretical support?

The answer is 'no,' and for interesting reasons. The case of 'think that' is not too hard to explain. It is well known that intuitions regarding 'think that' constructions and related propositional attitude verbs don't always seem to strictly track content. Examples abound, from Hesperus-Phosphorus cases to the pseudo *de re*. Our favoured view of judgements regarding 'think that' is that they are sensitive to lexical items and how closely they track the mental language of the subject of the attitude report. So it is not surprising that pejorative terms, if they differ in register, would indicate features of how the thinker is thinking the thought rather than strictly the content of the thought, impacting in the truth values of 'think that' sentences. Likewise, although in a less structured way, 'crude' and immature' in the cases above are meant to help bring out the means by which the thinker is thinking the thought. Ascriptions of properties like 'being immature' or 'being crude' not only emerge from how we act, but also from how we think our thoughts. These are not the only words to do so: in Hom's cases, for example, adjectives like 'racist' and 'anti-semitic' etc. seem to play the same role. After all, these properties are ascribed to people in virtue not only of what they think or feel about other people, but also in virtue of the way they think about them. In any case, neither of these constructions -'think that', and phrases containing expressions that bring out how thoughts are thought- are a serious threat to compositionally.

There is, of course, a chance that we are wrong and the semantics is strictly insensitive to these features. Fair enough, but all we really have to be right about is the claim that our interpretive intuitions are guided by these considerations, even if wrongly. Our dilemma for theorists like Hom is that if we are right his cases are not dispositive, and if we are wrong his cases

are ones in which we have defective intuitions. In other words, short of further argumentation that there is something special about GPTs substitution, we win either way.

7. Conclusion

Our goal in this paper was to debunk one of the most compelling arguments for the conclusion that slurs encode derogatory content as a matter of their lexical semantics, namely, substitutional arguments. Our method was to show that in general, words of different register (such as slurs but also non-slurring examples) can result in shifts in intuitive truth values in substitution cases like the ones used to argue for a semantic account of slur terms. Since non-slurring terms that differ in non semantic social features, like 'cutpurse' and 'pickpocket,' are highly plausibly semantically equivalent, we argued that substitutional arguments don't seem to elicit reliable intuitions in the case of slurs which also different in register from their non-slurring counterparts. We thus hope that we have cleared one of the main obstacles to research projects that take the derogatory content of slurs to be explained in non-semantic content manners. The argument, of course, does not refute semantic approaches to GPTs. But it should make them seem less explanatorily inevitable.

One final point; if we are wrong from the get-go and terms like 'cutpurse' and 'pickpocket' are not semantically equivalent as shown by the substitutional argument, we feel that the paper is still a win for us (albeit not one we expected). Given that the substitutional arguments generalize fairly well to most socially loaded terms, it would be a striking find that differences in register, sociolect, dialect or chronicity are associated with differences in semantic value. We are highly dubious, however, that this is the right conclusion to take from the arguments presented above.

References

Aristotle. Sophistical Refutations (179a33–b32). Several editions
Camp, E., 2013. Slurring perspectives. Anal. Philos. 54 (3), 330–349.
Copp, D., Sennet, A., 2015. What kind of a mistake is it to use a slur? Phil. Stud. 172 (4), 1079–1104.
Copp, D., Sennet, A., 2017. Pejoratives and ways of thinking. Anal. Philos. 58 (3), 248–271.
Crimmins, M., Perry, J., 1989. The prince and the phone booth: reporting puzzling beliefs. J. Philos. 86, 685–711.
Croom, A.M., 2011. Slurs. Lang. Sci. 2011 (33), 343–358.
Croom, A.M., 2013. How to do things with slurs: studies in the way of derogatory words. Lang. Commun. 33 (3), 177–204.
Diaz-Legaspe, J., Liu, Ch, Stainton, R., 2019. Slurs and register: a case study in meaning pluralism. Mind Lang. 34.
Fischer, E., Engelhardt, P.E., 2016. Intuitions' linguistic sources: stereotypes, intuitions and illusions. Mind Lang. 31 (6), 67–103.
Fischer, E., Engelhardt, P.E., 2020. Lingering stereotypes: salience bias in philosophical argument. Mind Lang. 35, 415–439.
Frege, G., 1892. Über Sinn und Bedeutung. Z. Philos. Philos. Krit. 100 (1), 25–50.
Halliday, M.A.K., 1973. Explorations in the Functions of Language. Edward Arnold, London.
Halliday, M.A.K., Matthiessen, C.M.I.M., 2004. An Introduction to Functional Grammar, third ed. Arnold, London.
Hasan, R., 2004. Analysing discursive variation. In: Young, L., Harrison, C. (Eds.), Systemic Functional Linguistics and Critical Discourse Analysis: Studies in Social Change. Continuum, London and New York, pp. 15–52.
Hom, C., 2008. The semantics of racial epithets. J. Philos. 105 (8), 416–440.
Hom, C., 2010. Pejoratives. Philos. Compass 5 (2), 164–185.
Hom, C., 2012. A puzzle about pejoratives. Phil. Stud. 159 (3), 383–405.
Hom, C., May, R., 2013. Moral and semantic innocence. Anal. Philos. 54 (3), 293–313.
Hom, C., May, R., 2018. Pejoratives as fiction. In: Sosa, D. (Ed.), Bad Words. Oxford University Press, Oxford.
Horn, L., 1985. Metalinguistic negation and pragmatic ambiguity. Language 61 (1), 121–174 (54 pages).
Jeshion, R., 2013. Expressivism and the offensiveness of slurs. Phil. Perspect. 27 (1), 231–259.
Predelli, S., 2003. Scare quotes and their relation to other semantic issues. Ling. Philos. 26 (1), 1–28.
Predelli, S., 2013. Meaning without Truth. Oxford University Press, Oxford.
Nunberg, G., 2018. The social life of slurs. In: Fogal, D., Harris, D., Moss, M. (Eds.), New Work on Speech Acts. Oxford University Press, Oxford.
Salmon, N., 1986. Frege's Puzzle. MIT Pres, Cambridge.
Salmon, N., 1989. Illogical Belief. In: Philosophical Perspectives 3: Philosophy of Mind and Action Theory. Ridgeview, Atascadero CA, pp. 243–285.
Soames, S., 1987. Substitutivity. In: Thomson, J.J. (Ed.), On Being and Saving: Essays in Honor of Richard Cartwright. MIT Press, Cambirdge, pp. 99–131.
Soames, S., 1988. Direct reference, propositional attitudes, and semantic content. In: Salmon, N., Soames, S. (Eds.), Propositions and Attitudes. Oxford University Press, Oxford, pp. 197–239.
Whiting, D., 2013. It's not what you said, it's the way you said it: slurs and conventional implicatures. Analytic Philosophy 54 (3), 364–377.
Williamson, T., 2009. Reference, inference and the semantics of pejoratives. In: Almog, J., Leonardi, P. (Eds.), The Philosophy of David Kaplan. Oxford University Press, Oxford, pp. 137–159.
Williamson, T., 2010. The use of pejoratives. In: Whiting, D. (Ed.), The Later Wittgenstein on Language. Palgrave, Basingstoke.