The silence of the slurs: inferences about prejudice under ellipsis*

Maria Esipova (University of Konstanz)

February 3, 2023

Abstract In this paper, I aim to establish whether the prejudice component of slurs can be preserved under ellipsis, with the ultimate goal of elucidating the semantic nature of this meaning component. To do so, I conducted an inference judgement experiment, asking the participants to assess how likely it is that a given speaker is prejudiced against a certain group based on how they respond to an antecedent utterance containing a slur targeting this group. As expected, responses in which the speaker repeats the slur give rise to the highest prejudice likelihood ratings, and responses in which the speaker indirectly challenges the antecedent use of the slur by replacing it with a neutral counterpart give rise to the lowest prejudice likelihood ratings. Crucially, elliptical responses in which the slur itself is not uttered, but that require lexical identity for the slur (*one* replacement targeting an antecedent noun slur) still yield higher prejudice likelihood ratings than elliptical responses that do not necessarily require lexical identity for the slur (bare particle and VP ellipsis responses to utterances containing a noun slur). These results, thus, call for a hybrid analysis of the prejudice component of slurs, whereby the act of uttering a slur gives rise to a performative prejudice-signaling effect, which necessarily gets lost under ellipsis, but slurs also have a non-performative truth-conditional (albeit not-at-issue) prejudice component, which gets preserved under ellipsis if the slur itself is recovered.

1. Introduction

Different types of content behave differently under ellipsis (see, e.g., Esipova 2019 and references therein for an overview).[1] For instance, certain presuppositions cannot be ignored during ellipsis resolution. This is true, for example, for presuppositions of predicates denoting various stages of events (*start*, *stop*, *continue*, etc.), as shown in (1a), as well as for presuppositions of factive predicates (*know*, *regret*, *be delighted that*, etc.), as shown in (1b) and (1c).[2]

*This paper is accompanied by a set of supplemental materials that includes the experiment files to be run on PCIbex Farm (Zehr & Schwarz 2018), a spreadsheet with the raw data, and the R code for the data analysis: https://osf.io/ys9qz/ Full thanks to follow.

[1]Throughout the paper, I will use the term *ellipsis* maximally broadly, encompassing both phenomena involving complete lack of phonological content as well as instances of substitution such as *do so* and *one* replacement in English.

[2]Modulo local accommodation (see Heim 1983 et seq.), in which case this presupposition still isn't ignored, but is rather treated as part of the at-issue content.

(1) *Presuppositions of some items (predicates denoting stages of events, factives, etc.)*

a. Pam stopped smoking, {but Kim didn't / and Kim did, too / and so did Kim}.
   (i) → Pam used to smoke.
   (ii) → Kim used to smoke.

b. Zoe knows that she is in danger, {and Ash does, too / and so does Ash}, (#although Ash is not actually in danger, they just believe that they are). (sloppy reading)
   (i) → Zoe is in danger.
   (ii) → Ash is in danger.[3]

c. Lea regrets leaving, {but Mia doesn't / and Mia does, too / and so does Mia}.
   (i) → Lea left.
   (ii) → Mia left.

On the other end of the spectrum are expressive contributions of items like *fucking*, *damn*, etc., which always get ignored during ellipsis resolution, as shown in (2).

(2) *Expressive component of items like 'fucking', 'damn', etc.*

A: Did you bring a fucking gun to my house?
B: No, I didn't. / Yes, I did. / Yes, I did so. / Yes, I brought one.
   (i) → A is experiencing strong emotions.
   (ii) ↛ B is experiencing strong emotions.

While these two cases appear to be categorical, this paper will focus on the much less clear-cut case of *slurs*, which are expressions that have a *denotational component* (they typically denote a group of individuals characterized by some socially relevant property such as race, gender, age, etc.) and a *prejudice component* (they convey some form of negative attitude towards this group because of this socially relevant property). More specifically, this paper will be addressing the question of whether—or rather, to what extent—the prejudice component of slurs can be preserved under ellipsis. To do so, we will be looking at exchanges like (3) and asking what we can infer about one's attitudes towards a certain group based on how they respond to antecedent utterances that contain slurs targeting this group.

(3) *Context: In the 'Blade Runner' universe, 'skinjob' is a slur for synthetic humans; the neutral term for them is 'replicant'.*

A: Did you see a skinjob?
B: No, I didn't. / Yes, I did. / Yes, I saw one.

Question probing the presence (+ strength) of the "prejudice inference":
Is B prejudiced against replicants? (+ How likely? / How prejudiced?)

Note that in the semantics literature (including much of the literature on slurs), it is common to ask the question about the presence of a certain inference in a certain environment, including elliptical environments, in a binary categorical fashion: we either get an inference

[3]This is a particularly interesting example, because under standard assumptions, *know that p* and *believe that p* are supposed to be truth-conditionally equivalent in their local contexts, yet the the elliptical conjunct cannot be interpreted as 'Ash also believes that they are in danger'.

or we don't. But, of course, this question can be asked in a gradient fashion to probe the strength of the relevant inference—and in cases when multiple factors might be affecting the target inference, we should be asking this question in a gradient fashion. I maintain that this is the case for what I will label the *prejudice inference* of slurs. There are at least two ways in which we can think about this inference in a gradient fashion: in terms of how likely it is that one is prejudiced against the targeted group (prejudice likelihood), or in terms of how prejudiced they are (prejudice intensity); this paper will focus on the former.

The main empirical question this paper aims to address is what factors affect the strength of the prejudice inference in various types of elliptical and non-elliptical responses to antecedent utterances containing slurs. And the main theoretical question is what this empirical picture tells us about the semantic nature of the prejudice component of slurs. I put forward a specific hypothesis about the interplay of semantic, pragmatic, and syntactic factors that affect said empirical picture (section 2) and tested this hypothesis experimentally via an inference judgement task (sections 3 and 4). I conclude, based on the results of this experiment, that the prejudice component of slurs is *partially* preserved under ellipsis in cases when the slur itself is necessarily recovered and, therefore, cannot be fully assimilated either to presuppositions of items like *stop* or to expressive contributions of items like *fucking*. Instead, it requires a hybrid analysis, whereby the act of uttering a slur gives rise to a performative prejudice-signaling effect, which necessarily gets lost under ellipsis, but slurs also have a non-performative truth-conditional (albeit not-at-issue) prejudice component, which gets preserved under ellipsis if the slur itself is recovered (section 5). I wrap up the paper by discussing the limitations of the present study and potential directions for future research (section 6).

2. Background

2.1. More on different types of content under ellipsis

Let us look again at the contrast between presuppositions of items like *stop* or *know* and expressives:

(4) *Presuppositions of some items (predicates denoting stages of events, factives, etc.)*
   a. Pam stopped smoking, {but Kim didn't / and Kim did, too / and so did Kim}.
      (i) → Pam used to smoke.
      (ii) → Kim used to smoke.
   b. Zoe knows that she is in danger, {and Ash does, too / and so does Ash}, (#although Ash is not actually in danger, they just believe that they are). (sloppy reading)
      (i) → Zoe is in danger.
      (ii) → Ash is in danger.
   c. Lea regrets leaving, {but Mia doesn't / and Mia does, too / and so does Mia}.
      (i) → Lea left.
      (ii) → Mia left.

(5) *Expressive component of items like 'fucking', 'damn', etc.*
   A: Did you bring a fucking gun to my house?

B: No, I didn't. / Yes, I did. / Yes, I did so. / Yes, I brought one.
(i) $\rightarrow$ A is experiencing strong emotions.
(ii) $\not\rightarrow$ B is experiencing strong emotions.

Now, there are many ways in which these two cases come apart.

First, there are syntactic differences. The items that trigger the presuppositions in (4) (*stop*, *regret*, *know*) are "the lexical verbs" of the vP targeted by the ellipsis. For concreteness, in this paper, I am assuming a non-lexicalist framework like Distributed Morphology, and, furthermore, that VP ellipsis (VPE) and *do so* replacement both target vPs (following Merchant 2013), and that *one* replacement targets nPs (following Harley 2014). So, under this set of assumptions, being "the lexical verb" of the target vP in VPE or *do so* replacement means being the sister to the verbalizer v° head.[4]

In contrast, *fucking* in (5) is at best an adjunct[5] inside an nP that, in turn, is targeted by *one* or is inside the vP targeted by VPE/*do so* replacement. The reason why this matters is because contributions of truth-conditional,[6] but not-at-issue adjuncts can sometimes be ignored under ellipsis as well:[7]

(6) *Context: A loves all scotch; they are in a bar with B.*
A: Could you get me another delicious shot of Laphroaig?
B: OK, I {will / will do so / will get you one}. (But I don't know how you drink this stuff, though, it's disgusting.)
(i) $\rightarrow$ A considers all shots of Laphroaig delicious.
(ii) $\not\rightarrow$ B considers all (or any) shots of Laphroaig delicious.

Second, the presupposition of *stop* is a precondition for the at-issue content of the utterance in (4a) to even make sense. Similarly, regarding emotive predicates like *regret*, the attitude holder believing the content of the complement is a precondition for them having feelings about it.[8] However, this is not the case for the expressive contribution of *fucking*; in (5), the speaker's affect is orthogonal to the at-issue content of the question or the response. Now, note that this is also not necessarily the case for *know* either (assuming that the at-issue content of *know that p* is equivalent to *believe that p*, which is not necessarily correct). However, it appears that *not* being a precondition for the at-issue content of the utterance to make sense is at least a prerequisite for ignorability under ellipsis.

Third—and most importantly for the purposes of this paper—acts of producing expres-

[4]These assumptions are not crucial for the broad purposes of the paper, as far as I can tell, and the syntactic underpinnings of the theoretical discussion in this paper can be restated in a different framework, if one so desires.

[5]It is also possible that expressives like *fucking* are some sort of late merged elements or even that they aren't part of the syntactic structure of the host utterance at all and only integrate with the host utterance prosodically (see Esipova 2022 on the different ways in which expressive content can "parasitize" on larger utterances).

[6]Truth-based, propositional, or any other term one prefers for regular, non-performative content, at-issue or not. I will use the term *truth-conditional* in this paper.

[7]See Esipova 2019, 2021; Sailor & Colasanti 2020 for more details. See also Esipova 2022 on how truth-conditional evaluative modifiers like *delicious* differ from expressives like *fucking* from the perspective of semantic composition.

[8]See Schlenker 2021 on presuppositions as epistemic preconditions.

sives like *fucking* in (5) are *performative*. Now, the notion of performativity has been used in many different ways in and outside of linguistics. Here I simply mean that the speaker achieves their expressive goals by virtue of producing a certain form and cannot achieve these goals without producing said form. For instance, if one wants to relieve their internal tension caused by anger or frustration by swearing (or punching the wall for that matter), they have to actually swear (or punch the wall). But, of course, anaphoric phenomena like ellipsis hinge on not producing the antecedent form. Thus, all purely performative effects will necessarily be lost during ellipsis resolution.[9]

Now, a few notes are immediately in order with respect to performativity and ellipsis. First, given cases like (6), the entailment only goes in one direction: if a given piece of meaning is preserved under ellipsis, then it cannot be purely performative. The reverse isn't true: non-performative content can still be ignored under ellipsis. Second, anaphoric reference to forms—or to acts of producing them—is, of course, a possibility. Thus, we can refer to the performative effects associated with the act of producing a certain form without obtaining them. Thus, in (7), B refers to A's expressive act to essentially assert that they have the same mental state as A,[10] but they do not themselves obtain the tension-relieving benefits of the expressive act they are referring to. We should, thus, be cautious about the possibility of such "metalinguistic" interpretations when assessing whether a given piece of meaning is preserved under ellipsis.[11]

(7) A: Fuck! / *punches the wall*
B: Yep, {same / that}.

2.2. Research question

Now, slurs like in (3), repeated below as (8), cross-cut the distinctions above.

(8) *Context: In the 'Blade Runner' universe, 'skinjob' is a slur for synthetic humans; the neutral term for them is 'replicant'.*
A: Did you see a skinjob?
B: No, I didn't. / Yes, I did. / Yes, I saw one.
Question probing the presence (+ strength) of the "prejudice inference":
Is B prejudiced against replicants? (+ How likely? / How prejudiced?)

First, from the syntactic perspective, *skinjob* in (8) is "the lexical noun" of the target constituent in *one* replacement (but not in VPE); i.e., in this case, it is like the presupposition triggering verbs in (4) and unlike *fucking* in (5). Thus, at least in this case, if the prejudice component of the slur is lost during ellipsis resolution, it is not because it is contributed by a (not-at-issue) adjunct.

[9]For those who like the "use" vs. "mention" distinction, we can think of performative acts as use via mention, and ellipsis as use without mention. For content that is purely performative then, there is no use without mention.

[10]Note that this is regular truth-conditional meaning. We can also routinely "quote" (or more precisely, demonstrate) performative expression of affect to convey truth-conditional meaning; see Esipova 2022 for a more detailed discussion of this phenomenon.

[11]This is why in this paper I only look at VPE and *do so* replacement, but not *do that*, as I believe the latter might have a higher potential for such metalinguistic uses.

Second, the prejudice component of slurs is not a prerequisite for the at-issue content of A's question or B's response to make sense; i.e., in this respect, it is like the expressive contribution of *fucking* in (5) and unlike the presuppositions in (4). Thus, it meets the prerequisite for ignorability under ellipsis.

As for performativity, this is where the main ultimate question of this paper emerges. We know that acts of uttering slurs do give rise to strong performative effects tied to their prejudice component—even in the absence of any relevant intent on the speaker's part (e.g., if they produce a slur without knowing what it means). In fact, even words that resemble slurs can famously give rise to some automatic effects like this.[12] But is the prejudice component of slurs purely performative? Can it be preserved if the slur itself is not uttered, but is recovered during ellipsis resolution?

2.3. Prior literature

Saab (2020) looks at Spanish data involving dialogues with elliptical responses to antecedents that contain ethnic slurs, as well as other items, such as the informal verb *morfar* 'eat' (as opposed to the stylistically neutral *comer*):

(9) A: ¿A cuántos sudacas viste en la fiesta?
to how-many South Americans$_{\text{PEJ}}$ saw.2SG in the party
"How many South Americans$_{\text{PEJ}}$ did you see at the party?"
B: Vi a tres <sudacas>, pero podrías evitar ese modo de
saw.1SG to three <South American$_{\text{PEJ}}$> but could.2SG avoid that way of
hablar de los sudamericanos. Yo nunca hablo así de ellos.
speaking of the South American I never speak so of them
"I saw three, but you could avoid that way of speaking about South Americans. I never talk that way."

(Saab 2020, (43))

(10) A: Qué morfaste?
what ate.2SG.INFORMAL
'What did you eat$_{\text{INFORMAL}}$?'
B: Una pizza <?>, pero no tolero cuando hablás tan informalmente.
a pizza but not tolerate.1SG when speak.2SG so informally
Yo nunca lo hago.
I never it do
"A pizza. But I don't tolerate when you speak informally. I never do it."

(Saab 2020, (21))

Saab compares the exchanges above to versions thereof where the target items are repeated in the response, rendering the response contradictory:

(11) A: ¿A cuántos sudacas viste en la fiesta?
to how-many South Americans$_{\text{PEJ}}$ saw.2SG in the party

[12]See also Mandelbaum & Young 2022 arguing that even the form of slurs themselves might not be completely arbitrary.

B: #Vi a tres sudacas, pero podrías evitar ese modo de
saw.1SG to three South American$_{\text{PEJ}}$ but could.2SG avoid that way of
hablar de los sudamericanos. Yo nunca hablo así de ellos.
speaking of the South American I never speak so of them

(Saab 2020, (44))

(12) A: Qué morfaste?
what ate.2SG.INFORMAL

B: #Una pizza morfé, pero no tolero cuando hablás tan
a pizza ate.1SG.INFORMAL but not tolerate.1SG when speak.2SG so
informalmente. Yo nunca lo hago.
informally I never it do

(Saab 2020, (22))

Saab concludes that the prejudice component of slurs and the stylistic effect of words like *morfar* are ignored under ellipsis, stating that "ellipsis is an apt strategy to nullify the bias encoded in some lexical items". His theoretical take is then that slurs and their neutral counterparts as well as pairs like *morfar* and *comer* are different phonological realizations of the same root, with the meaning differences between the two holding only at PF. Using the notions adopted in this paper then, Saab's claim amounts to saying that the prejudice component of slurs is purely performative.[13]

A major problem with Saab's empirical claim, however, is that it is categorical, but the extent to which the "bias" of a given lexical item is preserved in a response to an antecedent utterance containing said item might very well be gradient and affected by a variety of factors. Furthermore, the way we probe the extent to which said bias is preserved matters as well. In particular, in the exchanges above, B simply denies talking in a certain way, which can be taken to mean that they simply avoid producing certain forms. But then the contrast between (9)/(10) and (11)/(12) is trivial: of course, B is not violating their stated principles when they don't say the offending word. This tells us nothing about whether there is no non-performative "bias" component in the target items above that is still preserved when said item is recovered during ellipsis resolution (and that can be further mitigated, for instance, by the speaker directly challenging their interlocutor on their use of the offending item—as is the case in all the exchanges above).

The bottom line is that we need a more fine-grained empirical picture, which this paper is seeking to provide.

2.4. Hypothesis and predictions

In order to get at this more fine-grained empirical picture, I looked at paradigms like (13).[14] The exchanges are all set in a fictional universe where humans co-exist with other races, such as centaurs, dwarves, elves, orcs, etc. None of the races poses any inherent danger to another, but the social dynamics between different races can be quite complicated, and some individuals harbor prejudice against certain races. The exchanges all happen in the context

[13]Saab's claim is, in fact, stronger than this, because one can still maintain that a slur and its neutral counterpart are distinct roots, but their meaning differences lie exclusively in the performative dimension.

[14]I only looked at positive responses to avoid metalinguistic negation interpretations.

of a criminal investigation. The detective is a human and is questioning different witnesses who are all humans as well. The detective is extremely prejudiced against all non-humans and is constantly using various slurs. In any given case, the witness might or might not share the detective's prejudice, but even if they are strongly opposed to using a given slur themselves, they do not feel in a position to challenge the detective openly.

(13) a. *Context: 'Tusky' is a slur for orcs.*
Detective: Did you see a tusky?
Witness: Yes. ('Bare') / Yes, I did. ('VPE') / Yes, I saw one. ('One') / Yes, I saw a tusky. ('Slur') / Yes, I saw an orc. ('Nonslur')

b. *Context: 'Tusky' is a slur for orcs. This slur can also be used as a verb meaning 'to crawl' (for any race), because orcs are stereotyped as living in caves and, thus, having to crawl through narrow spaces all the time. The detective is asking a question about a human.*
Detective: What happened next? Did he tusky under the table?
Witness: Yes. ('Bare') / Yes, he did. ('VPE') / Yes, he did so. ('So') / Yes, he tuskied under the table. ('Slur') / Yes, he crawled under the table. ('Nonslur')

Question: How likely do you think that this witness is prejudiced against orcs?

The question in (13) is probing *prejudice likelihood*. I put forward the following complex hypothesis about the factors that determine prejudice likelihood ratings in cases like (13):

(14) *Hypothesis about prejudice likelihood ratings*

a. Slurs do have performative effects, so prejudice likelihood ratings are expected to be highest when the witness utters the slur themselves ('Slur').

b. Prejudice likelihood ratings are expected to be lowest when the witness indirectly challenges the detective by using the neutral term instead ('Nonslur'), in an attempt to minimize complicity (see, e.g., Cepollaro 2020 and references therein on the notion of complicity).

c. The prejudice component of slurs is not exclusively performative. Thus, when the slur itself is obligatorily recovered during ellipsis resolution, prejudice likelihood ratings are expected to be higher than when it isn't. I furthermore make the following assumptions about the identity requirements under ellipsis that determine when the slur itself is obligatorily recovered:

(i) Abstract identity (in the sense of Harley 2014) is only required for the "main stem" of the constituent targeted by *one* replacement/VPE/*do so* replacement—i.e., the minimal structure (the root plus any derivational morphology), responsible for the core lexical meaning of the sister of the x° head of the target xP.[15] Here the assumption is that only this structure needs to be lexically recovered for resolving the relevant anaphor (*one* in *one* replacement and *do* in VPE and *do so* replacement), while some other form of identity (e.g., properly constrained form of truth-conditional

[15]For instance, in the case of the slurs used in this study, this structure is presumably more complex than just the root and includes the *-y* suffix.

identity) is sufficient for the entire constituent.[16] Thus:

1. For noun slurs, we only require lexical identity for the slur itself in *one* replacement ('One'), but not in VPE ('VPE') or bare particle ('Bare') responses.
2. For verb slurs, we only require lexical identity for the slur itself in VPE ('VPE') and *do so* replacement ('So'), but not in bare particle ('Bare') responses.

If all the components of the hypothesis above are correct, we expect the following empirical picture:

(15) *Predicted prejudice likelihood ratings (from lowest to highest)*

a. Nouns: 'Nonslur' < 'Bare'/'VPE' < 'One' < 'Slur'
b. Verbs: 'Nonslur' < 'Bare' < 'VPE'/'So' < 'Slur'

Now, going in, I had additional considerations regarding prejudice likelihood ratings in cases like (13).

First, I expected verb slurs to be potentially harder to judge than noun slurs (as no perfect counterparts of such slurs seem to exist in English[17]) and possibly less "offensive" overall (due to the less direct link between their meaning and the targeted group), so no direct comparison of noun and verb slurs was planned.

Second, I conjectured that for some people, providing shorter responses to antecedents with slurs might also be a strategy to minimize complicity (especially in a context where one might not feel comfortable to directly challenge their interlocutor on their use of slurs), which, if true, could introduce further gradient distinctions across the 'Bare' vs. 'VPE' vs. 'One/So' conditions and potentially obscure some of the contrasts predicted by the hypothesis. Anticipating a little bit: even if this is indeed a viable strategy for some people, it didn't seem to have an ostensible effect in my results.

[16] Alternatively, we could entertain the idea that all roots/stems are recovered in the syntax, but their specific syntactic role can result in further differences with respect to how much we can "negotiate" their identity in semantics or pragmatics. I will not pursue this idea here, however.

[17] Of course, in English any word can be verbed, including in a slur-ish way—for instance, using the phrase 'to X it up', where X is a word denoting a group, to evoke some property the speaker associates with this group. But in this case, there arguably is no obvious neutral counterpart. Another case that comes to mind is the Russian verb *vycyganivat'* 'beg', which contains the root *cygan*, i.e., the neutral word for Romani people in Russian. But here the root itself is not a slur, the slur-ish nature of the verb comes from its stereotype-based meaning. Elin McCready (p.c.) brought the verb *to n***** lip* ('to smoke a cigarette or drink a beverage in a way that makes the tip very moist') to my attention, which is probably the closest English equivalent to the fictional verb slurs I used in my experiment, but is still seemingly lacking a succinct neutral counterpart. This, of course, highlights the general issue with slurs denoting event predicates—which also applies to stylistic pairs like the Spanish *morfar* vs. *comer* mentioned above—that it is less obvious that we can safely assume denotational equivalence of the "charged" and the neutral items in this case. Of course, there are also instances where denotational equivalence can be disrupted for slurs denoting predicates of individuals as well (*She's an orc, but not a tusky!*), but there seems to be a higher inherent potential for such disruption in event predicates.

3. Methods

I designed an inference judgement task using items like in (13) that involved 4 race–slur pairs (centaur–hoofy, dwarf–stunty, elf–leafy, orc–tusky). There were 10 experimental conditions, or, more precisely, two separate experiments (for noun slurs and for verb slurs) with 5 response types in each. Thus, there were 40 experimental items in total. Each participant saw 2 randomly selected trials per condition and 2 attention checks (22 trials total). The instructions described the universe and the general context of the exchanges (including the explicit caveat that the witnesses did not feel comfortable to openly challenge the detective on their use of the slur). The prejudice likelihood was assessed by dragging a slider on a pseudo-continuous scale (mapped to 0–100) from 'Not at all likely' to 'Very likely' (set by default to the middle of the scale). Participants were native speakers of English, recruited on Prolific (final $N = 128$, after excluding participants that failed the attention checks according to the a priori decided failure criteria) and paid £1.25 for completing the task. Figure 1 shows a typical trial.

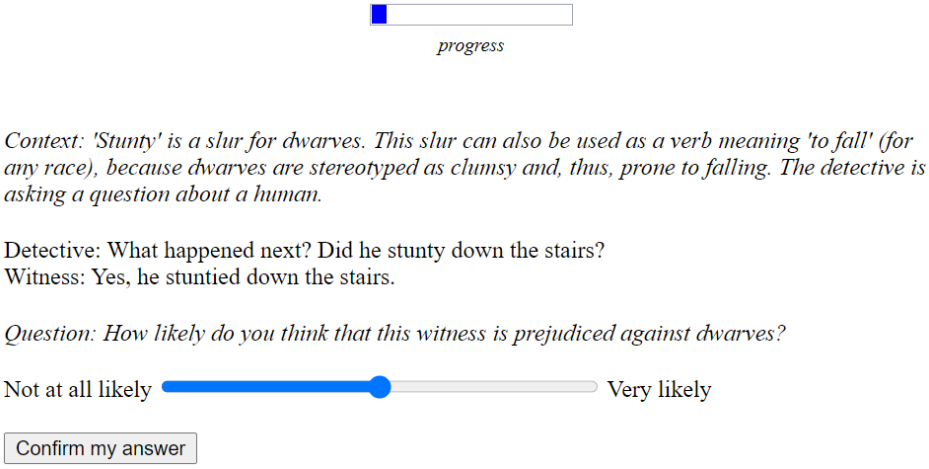

Figure 1: A typical trial.

A zip file containing the full experiment (the list of items, the instructions, the anonymized consent form, and the code for running the experiment on PCIbex Farm) is available as part of the supplemental package accompanying this paper (https://osf.io/ys9qz/).

4. Results

A spreadsheet with slightly pre-processed data (tidied and without Prolific IDs) is available as part of the supplemental package accompanying this paper (https://osf.io/ys9qz/). All statistical tests and plots were done in R (2022); the code is available as part of the same supplemental package.

The data were subsetted based on the part of speech (noun slurs and verb slurs), and a mixed effects linear regression model was run for each subset using the 'lmer' package, with the prejudice likelihood rating as the dependent variable, the response type as the independent variable, and a random intercept for participants. Planned comparisons between

specific pairs of response types were performed for these models using the 'emmeans' package (p-value adjustment method: Holm-Bonferroni). The results are visualized in Figure 2, and the statistics are given in Table 1.

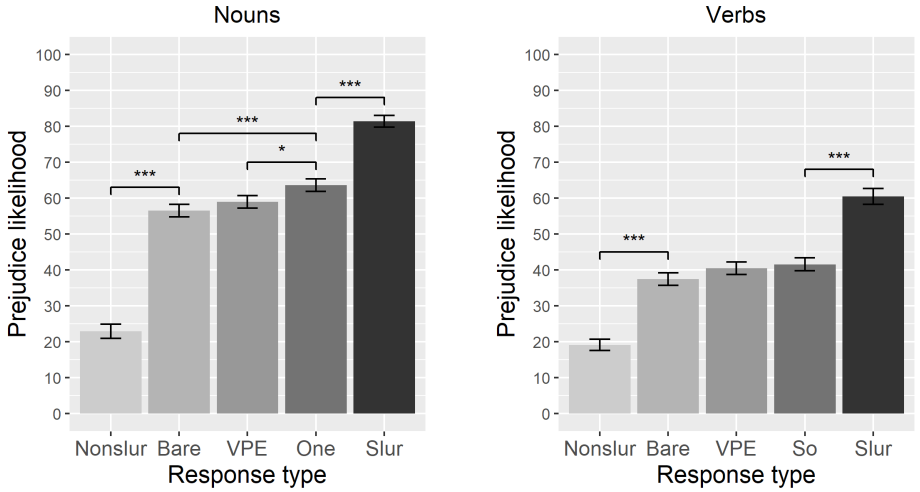

Figure 2: Bar charts showing mean prejudice likelihood ratings of different types of responses to antecedent utterances with noun and verb slurs, with key significant contrasts indicated. Error bars show standard error.

Table 1: Statistics for prejudice likelihood ratings for different types of responses to antecedent utterances with noun and verb slurs.

	Mean prejudice likelihood ± SE					**Planned contrasts (t values and adjusted p values)**				
	Nonslur	Bare	VPE	One	Slur	Nonslur/Bare	Bare/VPE	VPE/One	Bare/One	One/Slur
Noun slurs	22.9 ±2	56.5 ±1.73	58.9 ±1.74	63.6 ±1.73	81.4 ±1.65	*** $t = -17.552$ $p < .0001$	NS $t = -1.246$ $p = .2130$	* $t = -2.449$ $p = .0289$	*** $t = -3.695$ $p = .0007$	*** $t = -9.262$ $p < .0001$
	Nonslur	Bare	VPE	So	Slur	Nonslur/Bare	Bare/VPE	VPE/So	Bare/So	So/Slur
Verb slurs	19.1 ±1.57	37.4 ±1.75	40.5 ±1.74	41.5 ±1.79	60.4 ±2.24	*** $t = -10.133$ $p < .0001$	NS $t = -1.694$ $p = .1811$	NS $t = -.567$ $p = .5710$	NS $t = -2.261$ $p = .0719$	*** $t = -10.444$ $p < .0001$

As one can see, the results for noun slurs fully match the predictions in (15a). There is an ostensible contrast between indirectly challenging the slur by replacing it with a neutral term (lowest ratings) vs. repeating the slur (highest ratings) vs. providing an elliptical response (in-between ratings), as predicted by parts (14a) and (14b) of the hypothesis. Crucially for the purposes of this paper, as predicted by part (14c) of the hypothesis, there is a further (small, but statistically significant) contrast between cases when the slur itself is not necessarily recovered (bare particle and VPE responses) and cases when it is (*one* replacement), with the latter giving rise to higher prejudice likelihood ratings than the former. There was no statistically significant contrast between bare particle and VPE responses, suggesting that my concern that shorter responses might be a strategy to minimize complicity was unwarranted (or at least that this was not enough to yield any significant contrasts).

The results for verb slurs partially match the predictions in (14b). The three-way contrast between replacing the slur vs. repeating the slur vs. providing an elliptical response still holds, in line with parts (14a) and (14b) of the hypothesis. But there are no further statistically significant contrasts across the different types of elliptical responses—thus, failing to confirm part (14c) of the hypothesis for verb slurs.

I repeat the predictions and how the results compare to them in (16) and (17), respectively.

(16) *Predicted prejudice likelihood ratings (repeated)*

a. Nouns: 'Nonslur' < 'Bare'/'VPE' < 'One' < 'Slur'

b. Verbs: 'Nonslur' < 'Bare' < 'VPE'/'So' < 'Slur'

(17) *Statistically significant contrasts in prejudice likelihood ratings*

a. Nouns: 'Nonslur' < 'Bare'/'VPE' < 'One' < 'Slur' — full match

b. Verbs: 'Nonslur' < 'Bare'/'VPE'/'So' < 'Slur' — partial match

Additional post hoc models were run to see if there was any effect of the fictional race in the experimental items or any of the standard demographic factors (self-reported age, gender, race/ethnicity, country) on prejudice likelihood ratings, but didn't reveal any relevant contrasts.

5. Discussion

5.1. Discussion: noun slurs

The results for noun slurs suggest that the prejudice component of (noun) slurs is partially preserved when the slur itself is not uttered, but is necessarily recovered during ellipsis resolution (as is the case in *one* replacement responses in the current experiment)—in a way that goes beyond the effect of simply not challenging the antecedent slur (e.g., in bare particle and VPE responses).

Going back then to Saab's (2020) empirical claim evoked in subsection 2.3 that "ellipsis is an apt strategy to nullify the bias encoded in some lexical items", these results show that ellipsis does not "nullify" the prejudice component of slurs, but rather attenuates it. Theoretically then, it cannot be the case that slurs and their neutral counterparts are but different phonological realizations of the same root—or that the prejudice component of slurs is purely performative.

This calls for a hybrid analysis for the prejudice component of slurs that does not reduce it either to just a truth-conditional, but not-at-issue piece of content (be it a conventional implicature, as in Potts 2005, or a presupposition, as in Schlenker 2007) or to just a performative effect of producing a certain form on the context (as in Potts 2007 or Saab 2020).[18] Of course, such a hybrid analysis can only exist within a framework that does posit

[18]To be clear: here I am not talking about separating the denotational meaning of slurs from their attitudinal meaning—this goes without saying—but about the attitudinal component itself being of a mixed nature. Note also that here by *conventional implicature* I mean specifically conventional implicatures (CIs) in the style of Potts 2005. The systems in Potts 2005 and Potts 2007 are sometimes mixed up in subsequent literature, however, they are different on a fundamental level. Conventional implicatures in Potts 2005 are a special type of propositions; their type ends in a special, CI truth value. Also, Potts (2005) uses CIs to

a proper separation of the performative/expressive dimension from the non-performative (truth-conditional/propositional/etc.) one in the first place. Thus, Potts 2007 or a version thereof in Esipova 2022 that is intended to more explicitly target the performative vs. non-performative distinction are better equipped to handle hybrid entries for slurs than, for instance, Schlenker's (2007) response to Potts 2007, where expressive/performative contributions are reduced to a special type of presuppositions. My results, thus, should be viewed as another argument in favor of a proper separation of performative vs. non-performative meaning effects *on architectural grounds*, complementing, in particular, the compositionality-related architectural arguments in Esipova 2022.

Now, the point of this paper is to establish the need for a hybrid analysis for the prejudice component of slurs, not to propose a specific implementation of one, so I will remain neutral on the details of such an analysis (as well as on the specific formulation of the meaning effects themselves). For instance, one could maintain that the act of uttering a slur shifts the expressive dimension of the context (as in Potts 2007 or a version thereof in Esipova 2022[19])—this effect will be lost under ellipsis when the slur itself isn't uttered. However, in addition to that, the truth-conditional meaning of the slur includes a not-at-issue component[20] about the speaker's biased attitude—this component will be recovered during ellipsis resolution if the slur itself is recovered (i.e., when we require lexical identity for the slur).

5.2. Discussion: verb slurs

The results for verb slurs, however, do not provide support for the prejudice component of (verb) slurs being partially non-performative. I can think of two major reasons why the effect might have failed to obtain for verb slurs.

First, as anticipated prior to the experiment (see subsection 2.4), the prejudice likelihood ratings for verb slurs were lower than for noun slurs across the board—or more precisely, the range of the ratings was smaller, with the upper bound being substantially lower and all the contrasts therefore less pronounced (the range of the means across the response types was 22.9–81.4 for noun slurs and 19.1–60.4 for verb slurs). It is, thus, very plausible that the more subtle distinctions across the three types of elliptical responses were simply erased—especially in the context where the three-way contrast between challenging the slur, repeating the slur, and doing neither was much more salient (see also subsection 6.1).

Second, the relevant part of my hypothesis in (14c) was based on certain assumptions about identity requirements for different types of ellipsis, but, of course, identity requirements

analyze both supplements (appositives, sentence-level adverbs, parentheticals) and expressives, despite the former not being performative in the way expressives are. However, Potts (2007) departs from Potts 2005 in that he proposes a distinct treatment for expressives, one that is no longer truth-based: expressives now output new contexts. Contexts are modeled as tuples of several parameters, with expressives specifically altering the expressive parameter of the context, which consists of expressive indices, which in turn consist of individuals and numerical values within the $[-1, 1]$ interval aimed to capture a given individual's attitude towards another individual. See Esipova 2022 for a more detailed discussion of Potts 2005 vs. Potts 2007, as well as the remaining shortcomings of the latter for the purposes of capturing performative meaning effects.

[19]Of course, with the appropriate type of expressive indices added to the system, the need for which for slurs was already acknowledged in Potts 2007.

[20]Be it a presupposition, a conventional implicature, or yet another truth-conditional, but not-at-issue type of content—I remain agnostic as to which. In fact, I remain agnostic as to what the conceptually and empirically justified nomenclature of not-at-issue types of content should be.

on ellipsis resolution are themselves a complex and not yet fully understood phenomenon. It is, thus, entirely possible that while my assumptions about the identity requirements on *one* replacement vs. bare particle and VPE responses for noun slurs were correct, my assumptions about VPE and *do so* replacement vs. bare particle responses for verb slurs were not.

Thus, a natural follow-up would be a study that focuses just on the contrast across the different types of elliptical responses for verb slurs.

6. Conclusion

6.1. Limitations of the present study

Aside from addressing the main research question of this paper, the present study also aimed to establish a methodological precedent by (i) using a gradient scale in an inference judgement task about slurs, and (ii) using fictional slurs targeting fictional groups in such an experiment. Naturally, some of the methodological choices made when designing this study might have affected the results in various ways. Subsequent studies seeking to adopt a similar approach, thus, could make different methodological choices—in a way that is informed by the present study. In this subsection, I discuss some of these choices and speculate about their possible effect.

First and foremost, as mentioned before, the choice to (i) set all the exchanges in the experimental items in a context with a high power differential (a detective questioning witnesses), and (ii) draw the participants' attention to this differential by explicitly saying in the instructions that the witnesses would not feel comfortable openly challenging the detective on his use of slurs regardless of their own beliefs was deliberate. This was done with the intent of narrowing the space of potential responses to an antecedent utterance containing a slur (and, in particular, making sure that the 'Nonslur' responses, where the witness replaces the slur with a neutral term, would be the bottom baseline) and, therefore, maximizing the contrasts of interest. In hindsight, however, this appears to have made the contrast between indirectly challenging the slur, repeating the slur, and doing neither very salient while potentially minimizing the crucial contrasts across the different types of ellipsis—which non-linguist participants presumably already have less meta-awareness about.[21] Future studies seeking to employ similar methodology would need to consider how to balance these concerns. As far as the main research question of this paper goes, however, as I already mentioned in the previous section, now that the robust effect of replacing the slur vs. repeating the slur vs. doing neither has been established, follow-up studies could simply drop the slur-replacing and/or -repeating responses and focus just on the contrasts across the different types of ellipsis.

Second, as said at the outset, one can think of the strength of the prejudice inference triggered by slurs both in terms of prejudice likelihood and in terms of prejudice intensity. In the present study, I chose to go with the former, but subsequent studies could, of course, focus on the latter or even try to assess both. Impressionistically, however, it is not clear to me whether we can reliably disentangle the two, and, in fact, it is entirely possible that the

[21] This is evidenced, in particular, by the fact that none of the participants commented on the ellipsis-related contrasts in the optional field for general comments at the end of the experiment, while some did comment on the three-way contrast between replacing the slur, repeating the slur, and doing neither. See the data spreadsheet in the supplemental materials: https://osf.io/ys9qz/

participants in the present study were using the scale to cumulatively assess both prejudice likelihood and prejudice intensity when providing their responses.

Next, a common methodological point of contention is whether/when to use real-life slurs vs. fictional slurs in linguistic research. A common argument against using fictional slurs is that people might have harder time having intuitions about them. The problem with this reasoning, however, is that it makes an implicit assumption that everyone has the same intuitions about a given real-life slur, which ignores the diversity of backgrounds and experiences that even speakers of the same language might have when it comes to specific slurs (this, of course, is especially true for English). Note also that sometimes the same people who advocate for using real-life slurs in linguistic research end up using obsolete slurs or slurs that target groups that are assumed to be more privileged in their work in order to minimize "offensiveness" to the potential audience. This, to my mind, can create a much bigger methodological problem, as we may end up with judgements about "weak" slurs only, which can substantially skew our understanding of the empirical picture. Alternatively, we may have speakers mentally replace weaker slurs with stronger ones, if they don't have robust judgements about obsolete/rare/etc. slurs or even following an explicit instruction. But if the speakers have to do this mental replacement, this defeats the purpose of using real-life slurs in the first place. In fact, using fictional slurs in this case (and potentially explicitly encouraging the speakers to mentally replace them with real-life slurs *of their choice*) would level the ground for everyone, regardless of their experience with specific real-life slurs.

Most importantly, however, the present study demonstrates that it is, in fact, possible to get robust empirical effects with fictional slurs. If anything, this is an important methodological finding of this study, that people can exhibit clear, stable judgement patterns in offline tasks with fictional slurs (of course, fictional slurs would not be suitable for online processing tasks aiming to measure immediate reactions). Whether or not the participants of this study did the mental replacement when providing their judgements and, if they did, which real-life slurs they chose for this purpose is a black box—but I would argue that in the methodological context of this study this black box was a much better alternative to simply assuming that every English speaker will have the same experience with specific real-life slurs. Of course, more careful qualitative studies that would conduct sociolinguistic interviews and would take into account the backgrounds and experiences of individual participants could also be used to address the research questions raised in this paper (in which case I would, in fact, opt for using real-life slurs), but doing so with a large number of participants on Prolific doesn't seem feasible.

Finally, I would like to note that, upon looking at the individual judgement patterns, it appears that they were surprisingly homogeneous across the participants. Going in, I expected much more variation. In particular, there appear to be surprisingly few participants who simply did not associate using slurs with being prejudiced (and there was only one explicit comment to this effect in the optional general comments field). Upon reflection, however, this was likely a consequence of self-selection bias (setting aside the issue of how representative Prolific workers are in this respect to begin with). The premise that using slurs is an indicator of prejudice was built into the experiment, and people who strongly object to this premise would probably not choose to do the experiment in the first place. This is likely especially true in view of the fact that Prolific workers try to avoid having their submissions rejected by the researcher, as this affects their eligibility for future tasks (apart

from them wasting their time, of course), so if they believe (rightly or wrongly) that their submissions might be rejected on an "ideological" basis, they simply would not take on the task.[22] Thus, ironically enough, the power differential between researchers and workers on Prolific could have affected the representativeness of the results of this study.

6.2. Moving forward

Apart from direct follow-ups to the present study discussed above, there are multiple potential directions for expansion in future research. Below I mention just a few.

First, the methodology ground-tested in this paper can be extended to other types of not-at-issue attitudinal content (including the stylistic component of pairs like the Spanish *morfar* vs. *comer* mentioned in subsection 2.3), which have a tendency of being indiscriminately categorized and analyzed as "expressive" in the literature just because they are attitudinal and not-at-issue. Using the same methods as in this paper, we can then properly separate performative meaning effects of uttering a given form from truth-conditional (even if not-at-issue) attitudinal meaning contributions of a given item—which lends a solid architectural backbone to the expressive vs. non-expressive distinction.

Second, we can extend this methodology to other types of ellipsis in English and in other languages. Of course, an important relevant caveat is one that has already been brought up above: specific identity conditions for specific types of ellipsis are far from being comprehensively understood. In a way, studies like this one are combining two complex phenomena with multiple unknown variables within one equation in the hopes of learning something about some of those unknown variables.

Finally, in the fictional slurs used in this study, the prejudice meaning component intuitively comes from the combination of the root (evoking some stereotypical property of the targeted group) and the *-y* suffix. It would be interesting to look at the ellipsis-related behavior for affixes that productively create slurs from neutral terms for groups (fictional or real-life, if the latter exist) and, more generally, affixes that carry attitudinal meanings.[23]

References

Cepollaro, Bianca. 2020. *Slurs and thick terms: When language encodes values.* Lexington Books.

Esipova, Maria. 2019. *Composition and projection in speech and gesture*: New York University dissertation. URL https://ling.auf.net/lingbuzz/004676.

Esipova, Maria. 2021. On not-at-issueness in pictures. *Glossa: A Journal of General Linguistics* 6(1). 83. doi: 10.5334/gjgl.1314.

Esipova, Maria. 2022. Composure and composition. Ms., University of Konstanz, resubmitted. URL https://ling.auf.net/lingbuzz/005003.

Harley, Heidi. 2014. On the identity of roots. *Theoretical Linguistics* 40(3-4). 225–276. doi: doi:10.1515/tl-2014-0010.

[22] In the present study, all workers who finished the task were paid, even if their responses were excluded based on the attention checks, but, of course, the workers would not know that.

[23] See Esipova 2022 on the compositionality-related architectural differences between Russian diminutive suffixes used to performatively express affection and similar affect vs. derogatory and pejorative suffixes, which appear to contribute truth-conditional (even if not-at-issue) content.

Heim, Irene. 1983. On the projection problem for presuppositions. In Michael Barlow, Daniel Flickinger & Michael Wescoat (eds.), *Proceedings of the Second West Coast Conference on Formal Linguistics*, 114–125.

Mandelbaum, Eric & Steven Young. 2022. The sound of slurs: Bad sounds for bad words. To appear in *Oxford Studies in Experimental Philosophy*.

Merchant, Jason. 2013. Voice and ellipsis. *Linguistic Inquiry* 44(1). 77–108. doi: 10.1162/LING_a_00120.

Potts, Christopher. 2005. *The logic of conventional implicatures*. Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780199273829.001.0001.

Potts, Christopher. 2007. The expressive dimension. *Theoretical linguistics* 33(2). 165–198. doi: 10.1515/TL.2007.011.

R Core Team. 2022. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing Vienna, Austria. URL https://www.R-project.org/.

Saab, Andrés. 2020. On the locus of expressivity. Deriving parallel meaning dimensions from architectural considerations. In Eleonora Orlando & Andrés Saab (eds.), *Slurs and expressivity*, 17–44. Lexington Books.

Sailor, Craig & Valentina Colasanti. 2020. Co-speech gestures under ellipsis: a first look. Talk given at *The 94th Annual Meeting of the LSA*, New Orleans.

Schlenker, Philippe. 2007. Expressive presuppositions. *Theoretical Linguistics* 33(2). 237–245. doi: 10.1515/TL.2007.017.

Schlenker, Philippe. 2021. Triggering presuppositions. *Glossa: A Journal of General Linguistics* 6(1). 35. 1–28. doi: 10.5334/gjgl.1352.

Zehr, Jeremy & Florian Schwarz. 2018. PennController for internet based experiments (IBEX). doi: 10.17605/OSF.IO/MD832.