Slurs' Variability, Emotional Dimensions, and Game-Theoretic Pragmatics

Víctor Carranza-Pinedo[1,2][0000-0003-3253-4587]

[1] École Normale Supérieure de Paris, France
[2] Università degli Studi di Milano, Italy
victor.carranza@uni-muenster.de

Abstract. Slurs' meaning is highly unstable. A slurring utterance like 'Hey, F, where have you been?' (where F is a slur) may receive a wide array of interpretations depending on various contextual factors such as the speaker's social identity, their relationship to the target group, tone of voice, and more [16, 17, 32, 24, 12]. Standard semantic, pragmatic, and non-content theories of slurs have proposed different mechanisms to account for some or all types of variability observed, but without providing a unified framework that allows us to understand how different contextual factors simultaneously influence slurs' interpretation. To address this issue, I argue that slurs convey dimensional qualities such as, e.g., 'negative valence, neutral arousal, high dominance' instead of discrete emotional categories such as 'contempt'. Then, I translate this hypothesis into a game-theoretic model of slurs' interpretation inspired by Heather Burnett's pioneering work on identity construction [8, 9]. This new model, called 'Affective Meaning Games' (AMG), captures the variability of slurs and integrates pragmatic reasoning within an independently motivated psychological understanding of emotional states.

Keywords: Slurs, PAD Model of Emotions, Signaling Games, Expressivity

1 Introduction

Slurs are pejorative expressions employed to disparage individuals based on their association with social categories such as ethnicity (e.g., 'spic'), religion (e.g., 'kike'), gender (e.g., 'faggot'), etc. Even though slurs are typically used to express (and elicit) negative affective states such as contempt, hostility, or rage, it has been observed that their interpretation is not constant across different speech-act situations [16, 17, 32, 24, 12]. An utterance including a slur F like 'Hey, F, where have you been?' can receive a multitude of interpretations contingent upon various contextual factors including the speaker's social identity (e.g., their membership status in the group denoted by the slur), their relationship to the target group (e.g., whether they are close acquaintances), the intonation used (e.g. whether F is uttered with contempt or in a friendly tone), etc.

Semantic, pragmatic, and non-content theories of slurs have proposed diverse mechanisms to account for the phenomenon of variability. Firstly, theories that

semantically associate slurs with injurious attitudes explain slurs' variability by suggesting that some slurs are polysemous [16], or that perlocutionary mechanisms can influence the conventional expression of these attitudes [17]. Secondly, theories that view slurs as conversationally associated with clusters of negative stereotypes explain this variability by appealing to the role of speakers' communicative intentions and the varying degrees of severity of the associated stereotypes [7]. Lastly, theories that link slurs to prohibitions explain variability by claiming that these prohibitions can be enforced with varying degrees of severity or even 'suspended' in certain situations [1].

However, it is increasingly recognized that people infer agents' underlying emotional states by simultaneously integrating multiple contextual factors [15, 28, 29]. For instance, empirical evidence suggest that observers not only rely on facial expressions when inferring an agent's emotions but also consider other contextual cues such as body posture [3], background scenery [4], cultural norms [22], and so on, when these are available. Thus, in addition to evaluating the various mechanisms proposed to account for slurs' variability (i.e., meaning change, speaker's intentions, stereotypes, prohibitions, etc.), we also need a unified framework that allows us to understand both (i) how slurs' offensiveness arise and (ii) how different contextual factors interact with one another during the interpretation process, thereby giving rise to the attested variability.

To address (i), I characterize slurs as expressing values derived from continuous emotional dimensions (i.e., pleasure, arousal, and dominance) rather than discrete emotional categories (e.g., 'contempt'). Then, I argue that slurs' distinctive offensive profile is rooted in the high level of dominance they typically convey toward their target. To address (ii), I translate this hypothesis into a game-theoretic model of slurs' interpretation, drawing inspiration from Heather Burnett's pioneering work on identity construction [8, 9]. Under this new approach, called 'Affective Meaning Games' (AMG), a slur is indexically linked with a set of affective attributes (e.g., 'negative pleasure, neutral arousal, high dominance'), any one of which can emerge depending on prior assumptions about the speaker's emotional stance towards the target group. These assumptions are, in turn, influenced by different contextual factors (e.g., the speaker's identity, their relationship to the target, etc.), thereby providing a compact framework to analyze slurs' instability.

The paper is structured as follows. In Section 2, I distinguish two types of variability in slurs: first, variation in terms of the emotional states they express, and second, variation in terms of the offense they may provoke in others. Section 3 introduces the PAD model of emotions and proposes to characterize slurs' affective meaning in terms of negative pleasure and high dominance. Section 4 introduces Affective Meaning Games, and Section 5 explores some of their possible applications. Section 6 discusses further aspects of slurring speech acts and Section 7 concludes.

2 Two types of variation

The variation of slurs is Janus-faced. On the one hand, speakers use slurs to *express* a diverse spectrum of emotions. Typically, speakers display negative emotions such as contempt (i.e., 'Fs are not allowed here') or fear (i.e., 'Those Fs are invading us'). However, in different circumstances, such as when the slur is used among members of the target group, the speaker is typically characterized as expressing positive emotions like solidarity (e.g., 'Hey, my F, I have missed you') or pride (e.g., 'We should be proud of being Fs'). It is also noteworthy that the intensity of the emotion expressed can also vary, ranging from, e.g., mild condescension (e.g., 'I don't even see you as a F') to intense hatred (e.g., 'All Fs are greedy').

On the other hand, speakers typically *elicit* different degrees of offense among listeners by using slurs. This variation may occur in two ways: (i) across the lexical items employed or (ii) across the contexts in which slurs are uttered. With respect to (i), variation may occur across lexical items that target a single social group (e.g., 'beaner' may be considered more offensive than 'greaser') or different groups (e.g., 'chink' may be considered more offensive than 'guido'). With respect to (ii), variation may occur across different uses of the same expression, depending on who uses it (e.g., uses of 'faggot' within members of the LGTB community are considered less offensive than those performed by outsiders), the manner in which it is used (e.g., with a contemptuous or friendly intonation), etc.

Importantly, it has been observed that the offense a slur elicits is often orthogonal to the valence of the emotion it expresses. For example, although some members of the hippie subculture used the term 'spade' to express admiration or fondness for African Americans, the expression was still deemed offensive by the latter [27]. Conversely, certain insults that express extreme contempt or loathing are nonetheless perceived as inoffensive [32]. This occurs with insults directed at dominant groups (e.g., 'limey', 'toff', etc.) or that refer to an individual's personal traits (e.g., 'bastard', 'wimp', etc.). In contrast to these expressions, slurs distinctive scornful denigration is designed to manifest that the target is inferior [19].

These observations appear to undermine the idea that the offensiveness of slurs is related to the affective or any psychological states that the speaker is expressing through their use. For instance, Jeshion's [17] influential theory of slurs, which posits that all slurs are conventionally linked to the expression of contempt, has been criticized on the grounds that the expression of contempt is not sufficient nor necessary to explain the distinct scornful denigration that slurs inflict upon their targets [10]. An expression can be highly contemptuous without thereby being a slur, and a slur can express other emotions (e.g., fear, disgust, disdain or even amusement) and still be highly offensive.

Should we conclude that slurs' offensiveness is altogether independent of the emotions they convey? In what follows, I argue that this is not the case. To wit, emotions are not only analyzed as discrete emotional categories (e.g., 'contempt', 'joy', 'surprise', etc.), but also as states that we can characterize using basic affective dimensions (i.e., pleasure, arousal, and dominance). Thus, although the valence expressed by slurs may not be correlated to the offense they elicit, emotions have

other basic components that can assist in understanding this phenomenon. In the following section, I argue that slurs (i) tend to express affective states that qualify as negatively valenced but highly dominant, and (ii) that it is this latter dimension, dominance, that lies at the root of the offensiveness of slurs.

3 Slurs and the PAD model of emotions

What characteristics define an affective episode? Mehrabian and Russell [26] propose to describe affective episodes using a psychometric approach that employs three continuous, bipolar, and orthogonal dimensions: pleasure, arousal, and dominance. This approach, known as the 'PAD' model of emotions, was first introduced by Wundt [37] and is widely used for analyzing affective episodes in a continuous, rather than discrete, framework [34, 33, 25]. The three dimensions are defined as follows:

- PLEASURE: corresponds to a continuum that ranges from negatively valenced affective states (e.g., sadness) to positively valenced ones (e.g., joy) with respect to the stimulus. It is the evaluative component.
- AROUSAL: corresponds to the continuum ranging from low mental alertness (e.g., boredom) to high mental alertness (e.g., excitement). It is the physiological component.
- DOMINANCE: corresponds to the continuum ranging from the sensation of feeling controlled or submissive (e.g., frustration) to the sensation of feeling in control or powerful (e.g., anger) with respect to the stimulus. It is the relational component.

Dominance pertains to the degree to which an agent feels behaviorally constrained with respect to a stimulus (e.g., individuals, objects, events, etc.), on the basis of perceived qualities like physical strength, social status, hostility, etc. [26, 30]. Note that the level of dominance experienced is inversely proportional to the level of dominance perceived in the stimulus. For example, when provoked by stimuli perceived to be less dominant (e.g., an individual considered to be of 'lower status'), offenses are more likely to elicit anger than frustration (and vice-versa). With this being said, how can the content of slurs be characterized using the PAD dimensions?

1. Slurs typically express the speaker's negative appraisal of a specific group. For example, when the speaker utters 'That building is full of Fs', the listener is likely to infer that the speaker experiences DISPLEASURE with respect to F's target group or, similarly, that 'Fs are bad for being Fs'.
2. Interestingly, slurs do not seem to be significantly associated with a particular degree of AROUSAL. In contrast to other highly colloquial expressions such as 'fucking' or 'shitty', slurs don't come as infelicitous in situations where the speaker is only experiencing mild emotions. Slurs belong to the bigot's idiolect, rather than being reserved for extreme situations.
3. Lastly, slurs typically express that the speaker regards himself as superior to the target group. By uttering slurring sentences like 'Fs are not allowed here', the

speaker communicates that the members of the target group rank as low in worth, thereby attempting to establish a DOMINANCE hierarchy.

Why do slurs signal high dominance across different groups? One reason is that slurs are labels that arise as straightforward impositions to the target groups, and thus undermine their ability to build their own identity in an autonomous way [1]. Moreover, the act of slurring is 'action-engendering', that is, it grants permission for other forms of unjust treatment, such as physical or structural violence [21]. Therefore, slurs are uttered not only for the purposes of expressing a negative assessment, but also seek to establish or strengthen an unjust hierarchy between the speaker and the target through the expression of dominance (and, by extension, between the broader social groups to which they belong). I call this the 'valence-dominance' hypothesis.

It could be argued, however, that the expression of high dominance, which involves deeming individual targets as inferior, necessarily presupposes the expression of a negative evaluation. For example, in Jeshion's view [18, p. 133], slurs (i) express contempt towards a target group G, i.e., where contempt 'involves ranking another person as low in worth along the moral domain on a certain basis', and (ii) aim at specifying what members of G fundamentally are, based on the idea that belonging to G is a 'fundamental negative characteristic-defining feature of the targets'. In this theory, negatively evaluating G serves as the basis for treating its members as low in worth along the moral domain.

However, despite the fact that the aspects of high dominance and low pleasure often co-occur in slurring utterances, they are dissociated in many contexts. Indeed, it is possible to evaluate an individual negatively without expressing that they are inferior (e.g., when one qualifies someone as uninteresting or lethargic). Conversely, it is possible to express that an individual is lesser as a person without evaluating them negatively (e.g., racist ideologies about Chinese people are built on positive evaluations, such as that they are intelligent or hardworking). Being evaluated as good in some aspect doesn't preclude being simultaneously judged as inferior, and vice-versa.

How can the valence-dominance hypothesis explain slurs' offensiveness? As noted earlier, slurs can express emotions of opposite valence and still be regarded as offensive (e.g., contempt towards the target group or amusement at their expense). Nevertheless, in both cases, slurs invariably seek to dehumanize individuals by placing them beneath others within a dominance hierarchy. Hence, the valence-dominance hypothesis provides a straightforward account of slurs' offense: because slurs are linked with high-dominance states, their utterance warrants offense to those who find oppression detrimental to society. That is, their utterance provides moral justification for those who reject unjust forms of group-based hierarchy to take offense.

In this section, I have used the PAD model to characterize a subset of the vast array of affective states that can be potentially expressed by slurs in a particular utterance context. In the following, I will translate the valence-dominance hypothesis into a

probabilistic model of slurs' interpretation, in order to understand how slurs' variation emerges from the integration of multiple contextual cues.[1]

4 Affective Meaning Games

How can we operationalize emotions in a theory of meaning? Following Burnett's research on identity construction [8, 9], I postulate a structure ⟨Q, >⟩, where Q denotes a set of relevant affective qualities (e.g., high dominance or '[D+]') and > encodes relations of mutual exclusivity between them (e.g., individuals cannot experience a [P–] and [P+] state at the same time). As noted earlier, slurs don't correlate with a specific degree of arousal, so this dimension is omitted:

(1) Q = {[P+], [P–], [D–], [D+]}
    a. [P+] > [P–]
    b. [D–] > [D+]

Based on ⟨Q, >⟩, we can derive four distinct types of affective states α, such as the [P+, D–] state, labeled AFFILIATION, or the [P–, D+] state, labeled CONTEMPT. Importantly, these labels assemble different discrete emotional categories together. For example, CONTEMPT represents [P–, D+] states in general (e.g., rage, hostility, etc.), and not only contempt:

Table 1. Affective states α ∈ AFF

AFF	AFFILIATION	AMUSEMENT	ANXIETY	CONTEMPT
α	[P+, D–]	[P+, D+]	[P–, D–]	[P–, D+]

Then, I posit that for a given slurring term F, there is a non-slurring alternative F* that derogates the same social group G. For example, we assume that 'Spic' and 'Hispanic' are such alternatives, as the former probably emerged as a hypocoristic variant of the latter. Note, however, that we don't need to assume that F and F* are fully co-referential or etymologically related, but merely that they are salient lexical choices within the conversational interaction.

How can we characterize the link between the alternatives F/F* and the affective states α ∈ AFF that they have the potential to express? Since it is not possible to assign a stable interpretation to slurs across different contexts, I assume that the link between F/F* and affective states is *indexical*, that is, grounded on the statistical correlation

[1] In Section 2, I mentioned that slurs can express fear, despite fear being associated with low dominance behaviors such as freezing or fleeing. However, an alternative way to understand slurs is to view them as expressions of phobias, such as homophobia or xenophobia, which are affective dispositions based on the misrepresentation of the target as a threat to the agent's social privilege. Since phobias can trigger dominant behaviors, such as hostility or aggression, they warrant offense. Thanks to Isidora Stojanovic for drawing my attention to the non-dominant character of fear.

between the use of F/F* and a variety of affective qualities, any of which may be activated within a particular context [35, 13].

Specifically, I posit that slurs exhibit a stronger correlation with [D+] states, such as CONTEMPT, as opposed to [D–] states, such as AFFILIATION. To capture these regularities, I assign to F a probability distribution Pr(F|α), which represents the likelihood of uttering F given an affective state α [14]. Notably, the non-slurring alternative F* is associated with the distribution Pr(F*|α) = 1 – Pr(F|α):

Table 2. Affective-indexical meaning of F and F*

AFF	AFFILIATION	AMUSEMENT	ANXIETY	CONTEMPT
Pr(F|α)	0.3	0.6	0.4	0.7
Pr(F*|α)	0.7	0.4	0.6	0.3

Then, I assume that slurs are interpreted based on the listeners' L prior beliefs regarding the speaker's affective stance toward the group being targeted by the insult. Inspired by [8, 9], I represent L's prior beliefs as a probability distribution Pr(α), which denotes the probability distribution that the speaker S feels an affective state α towards target group G. In situations where L has no prior expectations about S's emotional stance towards G, we can represent Pr(α) as a uniform distribution over affective states:

Table 3. L's prior beliefs about S's affective stance α

AFF	AFFILIATION	AMUSEMENT	ANXIETY	CONTEMPT
Pr(α)	0.25	0.25	0.25	0.25

In other contexts, L's prior beliefs will be influenced by multiple factors, such as the speaker's identity, her relationship with the addressee, her previous actions, etc. For example, we can consider identities as social markers that we use to group each other and, more importantly, to shape our perception of their behavioral dispositions [2]. If someone is Catholic, we may assume that they experience [P+] states towards the Catholic church and endorse its teachings. If someone is Latino, we expect that they don't experience [D+] states toward Latinos, etc. While these assumptions may be proven incorrect, identities nevertheless guide our expectations regarding how others feel and behave.

Finally, once the speaker S utters a slur F directed at a social group G, L updates her prior beliefs by conditioning Pr(α) on F's affective meaning, Pr(F|α). In other terms, the interpretation process involves (i) combining the likelihood of F's signaling an affective state α with the L's prior beliefs about S's affective stance toward G, and then (ii) readjusting the outcome measure with a normalizing constant, i.e., the sum of these terms calculated for all affective states α ∈ AFF:

$$(2)\quad \Pr(\alpha|F) = \frac{\Pr(\alpha) \times \Pr(F|\alpha)}{\sum_{\alpha \in AFF} \Pr(\alpha) \times \Pr(F|\alpha)}$$

After introducing the fundamental elements of Affective Meaning Games, we can state its key conjecture: the affective information expressed by the use of a slur, perceived by a member of the audience, is constrained by the perceived affective relationship −according to that particular audience member− between the speaker and the social group G that is the target of the slur. In other terms, reasoning about S's potential emotions towards the target group G can alter the weighting of the various affective states $\alpha \in \text{AFF}$ in a particular context, thus giving rise to the variation observed in Section 2. In the following section, we will put this model into work by examining how it accounts for the Janus-faced variability of slurs' content.[2]

5 Explaining slurs' variability

In Section 2, it was observed that the impact of slurs varies across lexical items and different uses of the same lexical item. In this section, I explore how AMGs elucidate these phenomena and other aspects related to the usage and nature of slurs.

5.1 Variation across lexical items

The offensiveness of slurs varies across terms directed at the same group (e.g., 'beaner' vs. 'greaser') or different social groups (e.g., 'chink' vs. 'guido'). Our model explains this phenomenon as a result of the indexical character of the link between a given slur and values on the PAD dimensions. In contrast to conventional or conversational inferences, indexical associations are grounded in the co-occurrence of a particular sign and state, emerging from co-presence, causality, or other mechanisms [31, 5]. As a result, through repeated use and circulation, slurs gradually come to be associated with different indexical meanings that reflect the PAD values they regularly co-occur with. Thus, a term like 'chink' is more offensive than 'guido', or a term like 'beaner' more offensive than 'greaser', due to the higher intensity of the [D+] states or outcomes normally accompanying its use (e.g., hostile behaviors).

Indexicality also offers insight into how slurs' meaning shifts over time, for example during processes of appropriation [6]. Terms initially conveying a positive evaluation may be later reinterpreted as expressing dominance (e.g., 'redskin' or 'spade'), while terms expressing dominance are later linked to a positive evaluation (e.g., 'queer' or 'gay'). Our model attributes this phenomenon to the 'multilayered' character of indexical associations, which constantly acquire new meanings [36]. For example, using '-in' instead of '-ing' (e.g., 'fishin" rather than 'fishing') was seen as signaling 'casualness', but then it was also linked to an insincere or condescending persona [11]. As new indexical associations coexist with the old ones, interpreters

[2] From this point on, we may introduce further elaborations to the model. For instance, we could assume that speakers do not merely express their actual emotions through slurs, but also make strategic decisions about whether to employ them based on factors such as the social costs that result from their use, or whether the addressee is likely to approve their use [14]. I leave the exploration of these extensions to future research.

must be attentive to contextual factors during the inferential process, as we will see in the next subsection.

Finally, it is worth noting that indexicality can also account for the projective behavior of slurs, that is, the fact that slurs' content can survive entailment-canceling operators such as negations or disjunctions (e.g., the slur F in 'It is false that the building is full of F' elicits offense despite occurring under the syntactic scope of a negation). To wit, indexical associations are not restricted to lexical items, but rather apply to the phenomenon of variation between alternatives more generally. Any instance of human behavior, like clothing, habits, or activities, can index social (or affective) qualities, as long as they evoke a contrast between alternatives [13]. Therefore, since indexicality is not exclusively a linguistic phenomenon, entailment-canceling operators are ineffective in blocking slurs' expression of high dominance.

5.2 Variation across context of utterance

Slurs' offense varies with respect to multiple contextual factors such as the social identity of the speaker, their relationship with the addressee, their intonation, etc. We can illustrate how AMGs integrates these factors by describing four prototypical scenarios. In the first one, speaker S is not Latino and utters the sentence in (3), so we expect the slur to be interpreted as offensive by listener L.

(3) There will be a lot of spics at Mary's party.

To derive this interpretation, we assume that L lacks any preconception regarding S's emotional disposition towards Latinos. Hence, we plug the uniform distribution in Table 3 and the affective meaning of 'spic' (as outlined in Table 2) into the formula in (2). As a result, we derive that L is likely to interpret (3) as expressing the speaker's CONTEMPT (cf. the fourth row). Then, following the valence-dominance hypothesis, we explain why this utterance elicits offense:

Table 4. Neutral scenario

AFF	AFFILIATION	AMUSEMENT	ANXIETY	CONTEMPT
$Pr(\alpha)$	0.25	0.25	0.25	0.25
$Pr(spic\|\alpha)$	0.3	0.6	0.4	0.7
$Pr(\alpha)\cdot Pr(spic\|\alpha)$	0.075	0.150	0.100	0.175
$Pr(\alpha\|spic)$	0.15	0.30	0.20	0.35

Note that, regardless of whether S is perceived as evaluating Latinos positively or negatively by uttering (3), L's posterior beliefs will invariably tend to favor [D+] states. That is, if S is perceived as experiencing [P+] states (e.g., as uttering (3) in a friendly manner), he will be interpreted as expressing amusement at the expense of Latinos, implying that this social group is worthy of discriminatory practices such as racist jokes.

Secondly, in a scenario where S, despite not being Latino, possesses a certain 'insider' status within that community (e.g., S migrated to Latin America at a young age), we may expect (3) to express positive or negative emotions, but not necessarily to elicit offense. To account for this situation, we plug a distribution that favors [D–] states, and the affective meaning of the slur, in the formula in (2). As a result, we obtain that S is interpreted as expressing ANXIETY or AFFILIATION, which are inoffensive states:

Table 5. Insider scenario

AFF	AFFILIATION	AMUSEMENT	ANXIETY	CONTEMPT
Pr(α)	0.40	0.10	0.40	0.10
Pr(spic|α)	0.3	0.6	0.4	0.7
Pr(α)·Pr(spic|α)	0.12	0.06	0.16	0.07
Pr(α|spic)	0.29	0.15	0.39	0.17

In a third scenario, where S is Latino and utters (3), the slur will be even more clearly interpreted as inoffensive by L. The reason is that L will expect S to feel [P+] and [D–] states towards Latinos, as it is unlikely to feel members of one's groups as bad or worthy of contempt. As a result, L will interpret S as expressing AFFILIATION (e.g., affection, friendship, pride, etc.) towards members of the Latino community:

Table 6. Friendly scenario

AFF	AFFILIATION	AMUSEMENT	ANXIETY	CONTEMPT
Pr(α)	0.60	0.15	0.15	0.10
Pr(spic|α)	0.3	0.6	0.4	0.7
Pr(α)·Pr(spic|α)	0.18	0.09	0.06	0.07
Pr(α|spic)	0.450	0.225	0.150	0.175

Finally, what occurs when a speaker S uses a slur to dehumanize someone despite belonging to the group that is derogated? For example, a Latino may utter (3) with a contemptuous tone of voice, or use 'spic' in reference to one of their Latino employees, which some would interpret as offensive. In such a case, L may presume that S is trying to accommodate the presupposition that he doesn't identify as a Latino, as this appears necessary to interpret their utterance as an expression of contempt. However, if L rejects that presupposition, they may see S as experiencing ANXIETY instead, that is, as feeling unease at the mismatch between their culture of origin and the one they aspire to belong to.

In this section, we have examined how Affective Meaning Games explain the two types of variation discussed in Section 2. The first type is explained by the indexical nature of the association between slurs and affective dimensions. The second is explained by how such indexical associations are weighted against multiple background assumptions about the speaker's affective stance towards the target. We presented four scenarios to see the interplay between these assumptions. The first

scenario depicted typical instances of slurs being used as weapons. The second and third demonstrate how the harmful effects of slurs are diluted when the speaker belongs to, or is perceived as belonging to, the target group. Finally, the fourth scenario illustrates that in-group uses of slurs are not necessarily 'innocent', that is, don't always express positive emotions like pride or affiliation.

6 Comparison

In this section, I will compare Affective Meaning Games to two other models of slurs' interpretation that share a common interest in accounting for the unstable nature of slurs.

The first model, proposed by McCready and Davis [24] (based on [23]), uses different axiom schemas that interact with one another in order to infer the speaker's attitude towards the targeted group. These schemas represent our folk beliefs about the typical interpretation of an agents' emotions given certain assumptions (e.g., normative facts about the world, the agent's relationship to the target, etc.). The main feature of this model is its nonmonotonic quality, i.e., the possibility to override conclusions by adding more specific information to the set of premises [24, p. 265]. During the inference of slur's emotional/offensive impact, certain cues are thus considered more prevalent than others due to their specificity. For example, if the speaker uses a slur F, we will typically infer that he feels contempt towards members of the target group. However, if we also know that the speaker is a member of the target group, we will cancel the former inference and think instead that the speaker is expressing affiliation.

Hence, like our proposal, McCready examines the interpretation of slurs as part of a broader process of reasoning about the speakers' emotions from various contextual factors. In fact, McCready [23] notes that this proposal can also be modeled in terms of Bayesian reasoning, where we obtain various conclusions held with different probabilities instead of a single defeasible one [23, p. 259].

However, it is worth noting that interpreters typically weigh contextual factors (e.g., posture, social identity, facial expressions, etc.) based on their perceived reliability rather than their specificity [38, 20]. Indeed, co-occurring cues that are equally specific can be in conflict with each other (e.g., a positive facial expression can accompany a contemptuous tone of voice). As a result, interpreters are sometimes required to assess the cues' relative degree of reliability independently of how specific they are (cf. scenario four in section 5.2).

In the second model, proposed by Popa-Wyatt and Wyatt [32], slurs are conversational moves that subordinate individuals by assigning them a lower discursive role within a conversation game. Moreover, the discursive roles of the participants draw upon and reinforce long-standing social roles, thus creating or reinforcing an unjust power imbalance. As a result, this theory explains slurs' offensiveness by appealing (i) to the allocation of a dominant/submissive discursive role to the speaker and target, respectively, and (ii) to the social roles they evoke in the history of oppression. Importantly, for the offense to occur, the speaker must 'fit'

the oppressor's role; if the speaker cannot fit that role (e.g., because they belong to the oppressed group), then the assignment won't be felicitous and the offense can't take place.

Popa-Wyatt and Wyatt's theory accounts for many instances of slur's variation. For example, it predicts that 'chink' is more offensive than 'guido' due to the greater degree of oppression experienced by people of Chinese descent. Similarly, it explains that 'spic' is not offensive when uttered by a Latino because such person doesn't fit the oppressor role. However, the authors acknowledge that some cases cannot be explained by appealing to group-membership alone, and that 'other flags, such as tone, familiarity between the participants, appropriateness of context, will all modulate whether the felicity condition is met.' [32, p. 22]. The authors suggest that these conditions are additive, meaning that if enough of them are present, offense arises. Nevertheless, as mentioned before, contextual cues are not always in harmony with each other, such as when a member of the group derogated utters a slur with a contemptuous tone of voice. Therefore, adding these cues up may not be always adequate to interpret the slur. While group membership is usually a reliable cue, it interacts with other factors that have the potential to become more reliable during the interpretation process.

7 Conclusion

This paper has put forward a novel account of on the affective meaning of slurs, which posits that it is grounded in the valence and dominance dimensions of emotions. By focusing on the dominance dimension as the key factor driving slurs' offensive nature, this paper opens a new perspective on the complex interplay between emotions, communication, and power. The proposed game-theoretic model presents a comprehensive framework for analyzing the variability of slurs' affective and offensive impact, and its flexibility holds promise for extending the analysis to other forms of injurious expressions, such as pejorative nicknames and particularistic insults. While empirical testing is needed to confirm the model's predictions, the proposed framework constitutes a first step towards underlying the cognitive mechanisms that govern the interpretation of non-linguistic expressions, behaviors, and visual stimuli, thereby advancing our understanding of how affective information is transmitted within social interactions.

Acknowledgements. I would like to express my gratitude to Elin McCready, Heather Burnett, Michael Franke, Filippo Domaneschi, Daniel Gutzmann, and particularly to Elisa Paganini, Márta Abrusán and Isidora Stojanovic for their valuable feedback and suggestions on earlier versions of this paper. All errors or omissions that remain are mine.

References

1. Anderson, L., Lepore, E.: Slurring words. Noûs 47, 25–48 (2013). https://doi.org/10.1111/j.1468-0068.2010.00820.x
2. Appiah, K.A.: The Ethics of Identity. Princeton University Press, Princeton (2010). https://doi.org/10.1515/9781400826193
3. Aviezer, H., Trope, Y., Todorov, A.: Body cues, not facial expressions, discriminate between intense positive and negative emotions. Science 338(6111), 1225–1229 (2012). https://doi.org/10.1126/science.1224313
4. Barrett, L.F., Kensinger, E.A.: Context is routinely encoded during emotion perception. Psychological Science 21(4), 595–599 (2010). https://doi.org/10.1177/0956797610363547
5. Beltrama, A.: Social meaning in semantics and pragmatics. Language and Linguistics Compass 14(9), e12398 (2020). https://doi.org/10.1111/lnc3.12398
6. Bianchi, C.: Slurs and appropriation: An echoic account. Journal of Pragmatics 66, 35–44 (2014). https://doi.org/10.1016/j.pragma.2014.02.009
7. Bolinger, R.J.: The pragmatics of slurs. Nous 51(3), 439–462 (2015). https://doi.org/10.1111/nous.12090
8. Burnett, H.: Sociolinguistic interaction and identity construction: The view from game-theoretic pragmatics. Journal of Sociolinguistics 21(2), 238–271 (2017). https://doi.org/10.1111/josl.12229
9. Burnett, H.: Signalling games, sociolinguistic variation and the construction of style. Linguistics and Philosophy 42(5), 419–450 (2019). https://doi.org/10.1007/s10988-018-9254-y
10. Camp, E.: Slurring perspectives. Analytic Philosophy 54(3), 330–349 (2013). https://doi.org/10.1111/phib.12022
11. Campbell-Kibler, K.: Listener perceptions of sociolinguistic variables: The case of (ING). Ph.D. thesis, Stanford University, Stanford CA (2005)
12. Davis, C., McCready, E.: The instability of slurs. Grazer Philosophische Studien 97(1), 63–85 (2020). https://doi.org/10.1163/18756735-09701005
13. Eckert, P.: Variation and the indexical field. Journal of sociolinguistics 12(4), 453–476 (2008). https://doi.org/10.1111/j.1467-9841.2008.00374.x
14. Henderson, R., McCready, E.: Dogwhistles and the at-issue/non-at-issue distinction. In: Gutzmann, D., Turgay, K. (eds.) Secondary content, pp. 222–245. Brill (2019). https://doi.org/10.1163/9789004393127_010
15. Hess, U., Hareli, S.: The role of social context for the interpretation of emotional facial expressions. In: Mandal, M.K., Awasthi, A. (eds.) Under-standing facial expressions in communication, pp. 119–141. Springer (2015). https://doi.org/10.1007/978-81-322-1934-7_7
16. Hom, C.: The semantics of racial epithets. Journal of Philosophy 105(8), 416–440 (2008). https://doi.org/10.5840/jphil2008105834
17. Jeshion, R.: Expressivism and the offensiveness of slurs. Philosophical Perspectives 27(1), 231–259 (2013). https://doi.org/10.1111/phpe.12027
18. Jeshion, R.: Slur creation, bigotry formation: The power of expressivism. Phenomenology and Mind (11), 130–139 (2016). https://doi.org/10.13128/Phe_Mi-20113
19. Jeshion, R.: Varieties of pejoratives. In: Khoo, J., Sterkin, R. (eds.) Rout-ledge Handbook of Social and Political Philosophy of Language (2020). https://doi.org/10.4324/9781003164869

20. Kayyal, M., Widen, S., Russell, J.A.: Context is more powerful than we think: contextual cues override facial cues even for valence. Emotion 15(3), 287 (2015). https://doi.org/10.1037/emo0000032
21. Lynne, T.: Genocidal language games. In: Ishani Maitra, M.K.M. (ed.) Speech and Harm, pp. 174–221. Oxford University Press (2012). https://doi.org/10.1093/acprof:oso/9780199236282.003.0008
22. Masuda, T., Ellsworth, P.C., Mesquita, B., Leu, J., Tanida, S., Van de Veerdonk, E.: Placing the face in context: cultural differences in the perception of facial emotion. Journal of personality and social psychology 94(3), 365 (2008). https://doi.org/10.1037/0022-3514.94.3.365
23. McCready, E.: Emotive equilibria. Linguistics and Philosophy 35(3), 243–283 (2012). https://doi.org/10.1007/s10988-012-9118-9
24. McCready, E., Davis, C.: An invocational theory of slurs. Proceedings of LENLS 14 (2017)
25. Mehrabian, A.: Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. Current Psychology 14(4), 261–292 (1996). https://doi.org/10.1007/BF02686918
26. Mehrabian, A., Russell, J.A.: An approach to environmental psychology. MIT Press (1974)
27. Nunberg, G.: The social life of slurs. In: Daniel Fogal, Daniel W. Harris, M.M.(ed.) New Work on Speech Acts, pp. 237–295. Oxford University Press (2018). https://doi.org/10.1093/oso/9780198738831.001.0001
28. Ong, D.C., Zaki, J., Goodman, N.D.: Affective cognition: Exploring lay theories of emotion. Cognition 143, 141–162 (2015). https://doi.org/10.1016/j.cognition.2015.06.010
29. Ong, D.C., Zaki, J., Goodman, N.D.: Computational models of emotion inference in theory of mind: A review and roadmap. Topics in cognitive science 11(2), 338–357 (2019). https://doi.org/10.1111/tops.12371
30. Oosterhof, N.N., Todorov, A.: The functional basis of face evaluation. Proceedings of the National Academy of Sciences 105(32), 11087–11092 (2008). https://doi.org/10.1073/pnas.0805664105
31. Peirce, C.S.: Philosophical writings of Peirce, vol. 217. Courier Corporation (1955)
32. Popa-Wyatt, M., Wyatt, J.L.: Slurs, roles and power. Philosophical Studies 175(11), 2879–2906 (2017). https://doi.org/10.1007/s11098-017-0986-2
33. Reisenzein, R.: Pleasure-arousal theory and the intensity of emotions. Journal of personality and social psychology 67(3), 525 (1994). https://doi.org/10.1037/0022-3514.67.3.525
34. Russell, J.A.: A circumplex model of affect. Journal of personality and social psychology 39(6), 1161 (1980). https://doi.org/10.1017/S0954579405050340
35. Silverstein, M.: Shifters, linguistic categories, and cultural description. Meaning in anthropology pp. 11–55 (1976)
36. Silverstein, M.: Indexical order and the dialectics of sociolinguistic life. Language & communication 23(3-4), 193–229 (2003). https://doi.org/10.1016/S0271-5309(03)00013-2
37. Wundt, W.: Compendio de psicología. La España Moderna (1896)
38. Zaki, J.: Cue integration: A common framework for social cognition and physical perception. Perspectives on Psychological Science 8(3), 296–312 (2013). https://doi.org/10.1177/1745691613475454