The expression of hate in hate speech
Teresa Marques
University of Barcelona

Abstract

In this paper, I claim that hate speech expresses hate. I also answer some objections to expressivist views. First, I will offer a model of expressively normative speech. I argue that there is speech that is expressive of an emotion or sentiment, by presupposing it. This presupposition can become common ground by enacting the normative appraisal that is constitutive of the emotion or sentiment expressed (if the utterance is felicitous). This supplements current pragmatic accounts of hate speech in terms of the illocutionary enactment of given norms that nonetheless do not explain what the source of those norms is. Current research on hate confirms the working hypothesis that it is an appraisal state. The appraisals that constitute hate sentiments, and that are presupposed in hate speech, can become common ground among a population. I will show that this view is not undermined by several objections to expressivist views, and that it does not face some of the limitations of other accounts. Views that rely on the existence of preexisting systems of discrimination to analyze hate speech can’t explain hate crimes fueled by hate speech against social groups that are not historically discriminated. The account I offer can.

Keywords: hate speech, expressives, pressupositions, hate sentiment, discrimination

This paper argues for the apparently trivial thesis that hate speech expresses hatred. Currently, this is not seen as trivial, and it is also not popular. And yet, as I argue, current work on semantics, pragmatics, and the emotions offers the tools for explaining speech that is expressive of emotions, in general, and of hate in particular.

Various authors claim that the opposite of what I claim is true. Jeremy Waldron considers that talking of “hate” is distracting, since the regulation of hate speech, which he favors, should not be the regulation of people’s thoughts or feelings (Waldron 2012, p. 34ff). On Waldron’s theory, hate speech is harmful because it undermines people’s assurance of dignity in society, i.e., it undermines people’s assurance that they will not face violence, aggression, discrimination or exclusion. Alexander Brown argues that it is a “myth” that emotions, feelings, or attitudes of hate or hatred are part of the essential
nature of hate speech. (Brown 2017a). His view is that the term “hate speech” is a family resemblance concept (Brown 2017b). But Katherine Gelber offers a conceptual understanding of hate speech that can be used to design policies to regulate “sufficiently harmful speech”, although the assumption that hate speech “implies the presence of an emotion of hate” is mistaken (Gelber 2019). She defers to Rae Langton’s distinction, after Austin, between constitutively harmful speech and causally harmful speech. However, Gelber considers that this does not (or need not?) “imply” an emotion of hate. She proposes instead “a systemic discrimination approach” to define a “narrowly construed category of ‘hate speech’ as speech that harms to a sufficient degree to warrant government regulation.”

These views illustrate two current lines of thought. On the one hand, there is widespread consensus against the idea that hate is essentially tied to hate speech is. On the other hand, there is no consensus about what hate speech is. I think that there are insightful and true aspects of each theory, but that those authors are mistaken in rejecting the idea that hate speech expresses hate. I agree with Waldron that hate speech can have the effect of undermining citizen’s assurance of dignity. I agree with Brown that there are different ways to understand what “hate speech” means, even if I’m skeptical of the theoretical usefulness of “family resemblances.” Finally, although I agree with a lot of Gelber’s view, I disagree that this precludes hate speech to be expressive of hate.

In this paper, I want to contribute to the “conceptual understanding” of hate speech as Gelber calls it. This means that I think that there is such a thing as hate speech, and that some conceptual understanding is possible. I am not however giving recommendations for policy making or for legal regulation. This paper is primarily one in philosophy of language, with some input from the philosophy and psychology of emotions. We can try to understand what hate speech is and how it works. This is a
necessary precondition for appreciating when it can reach the dangerous levels that require legal regulation, or when it is likely to lead to hate crimes, but also to understand when it does not require regulation.

This paper is organized as follows. In the first section, I summarize an expressive-normative account of pejorative discourse that was recently defended by Marques and García-Carpintero (2020). I suggest that it can be adapted to a general framework for expressive speech, and that it can accommodate hate speech. This is compatible with some aspects of pragmatic accounts of harmful speech. The main difference between the expressive-normative account of Marques and García-Carpintero, and the pragmatic theories of Langton (2012, 2018) or McGowan (2004, 2009, 2019), is that the former expressive-normative view identifies the enacted of norms as the normative appraisals that constitute the expressed attitude (emotions). Other pragmatic views do not easily distinguish between different kinds of speech-enacting norms. What is missing for the application of the framework to hate speech is, then, independent support for a working hypothesis – that hate is an attitude that is itself a normative appraisal state.

In the second section, I review some recent work on the philosophy and psychology of emotions that corroborates this hypothesis. The account thus requires a combination of conversational contexts as contexts that can be updated with the normative requirements of the speech acts performed, even those provided by affective states, and an account of emotions and sentiments themselves as normative or evaluative appraisals. The positive claim I advance, then, is that hate speech is illocutionarily expressive and presupposes ongoing hate as a sentiment which “organizes people’s social world and helps strengthening the connection of the ingroup (“ingroup love”) at the expense of various outgroups (“outgroup hate”)” (Fischer et al. 2018, p. 311).
In the final section of the paper, I reply to Brown’s (2017a) objections to the idea that hate speech has any necessary connection to hate as an emotion. This reply suffices for replying to other existing objections.

1 An expressive presuppositional account

In this section I introduce and then expand on what Marques and Garcia-Carpintero (2020) call a “presuppositional expressive-normative account” of pejoratives and slurs. This account is both presuppositional and expressive, and that means that an utterance that literally uses a slur presupposes an attitude of contempt, thereby expressing contempt. I first present the motivation for expressivist and for presuppositional views. I then introduce the idea of expressive-presuppositions in richer contexts, and finally I contrast this theory with pragmatic accounts of hate or harmful speech.

1.1 Other expressivist and other presuppositional theories – the case of pejorative language

Robin Jeshion (2013, 2018) offered an expressivist, use-conditional, view of derogatory language, along the lines of Kaplan (1999). She argues that pejorative discourse is semantically expressive of contempt, which she analyses as a moral emotion. In some of her writings, she argues that words that can change their meaning and become derogatory, thereby contributing to create bigotry (Jeshion 2016). Words become derogatory by coming to be conventionally associated with the expression of contempt. For Jeshion, understanding derogatory language requires understanding “the moral psychological structure of contempt” (Jeshion 2016: 133). Uses of a slur convey that its targets have lesser standing as humans, that they are underserving of the full respect that is owed to persons qua persons and thus that using a slur dehumanizes its targets (ibid.: 131). Since
the expression of contempt is semantically encoded in the meaning of a slur, i.e., it is a linguistic convention that the pejorative is expressive of contempt, the question arises as to whether the expression of contempt requires that the speaker actually feels contempt. Jeshion is not completely clear on this point (Jeshion 2018: 97).

Other authors, like Boisvert (2008), argue for hybrid accounts of evaluative language more generally, inspired by a similarly hybrid expressivist explanation of pejoratives. As Boisvert argues, evaluative utterances express both a cognitive content (a proposition that can be true or false) and a conative attitude (attitudes like approval, disapproval, dislike, contempt, etc.). On expressive-assertivism, if a speaker utters (correctly and literally) a sentence that contains an evaluative (ethical) predicate in an extensional context, the speaker performs both a direct expressive illocutionary act and a direct assertive illocutionary act, where the expressive performed expresses an attitude toward the things that have the property picked out by the ethical predicate. Crucially, the correct and literal utterance of a sentence can perform an assertive or an expressive speech act even if the speaker lacks the relevant mental states. A speaker can assert even if she does not believe what she asserts, and she can express contempt for certain people even if she does not feel it.

As several authors persuasively argue, pejoratives and slurs seem to presuppose their derogatory force. The argument is given on the basis of their behavior when embedded under negation, as antecedents of conditionals, embedded under modal operators, or in interrogative or directive mood. The idea that pejoratives and slurs conventionally presuppose the expression of a negative evaluation of, or contempt for, the group that is the target of the slur or pejorative is advocated for instance by Macià (2002, 2014), Schlenker (2007), Cepollaro (2016, 2020), or Cepollaro and Stojanovic (2016). There are however differences in theories about what is presupposed.
Consider for instance Philippe Schlenker’s (2007, 238; 2016) account of the meaning of slurs. For Schlenker, the expressive presupposition of a slur $S$ is:

The agent of the context believes in the world of the context that the group targeted by $S$ are despicable.

This is a clear-cut condition on a Stalnakerian context. For Stalnaker, to assert that $\varphi$ is to propose that $\varphi$ become common ground:

It is common ground that $\varphi$ in a group if all members accept (for the purpose of the conversation) that $\varphi$, and all believe that all accept that $\varphi$, and all believe that all believe that all accept that $\varphi$, etc. (Stalnaker, 2002, p. 716)

Acceptance is the relevant belief-like attitude targeted through assertion. It need not be a full belief; it might be an assumption, or a mere ‘acceptance for the purposes of an argument or inquiry’. For Stalnaker (1978), an accepted assertion comes to be presupposed afterwards. Now, linguistic presuppositions are conventionally associated with lexical items or constructions. To understand these constructions, it is necessary to understand the associated presuppositions. In the case of slurs, an audience that understands and accepts an utterance with a slur would updates its common ground with this presupposed proposition of the above form.

However, there are well-known objections to Schlenker’s account. Williamson (2009: 151-2), for instance, has argued that that presupposition does not capture the normative status of slurs. Take the example “honky”, which is a slur for poor white people in the USA. If a speaker says (1), he is asserting that the demonstrated person likes Fox News, and is, according to Schlenker, presupposing that he himself believes that poor white people like are despicable.

(1) That honky likes Fox News.

If this is what the speaker communicates and what the audience accepts, then we have left out something fundamental about the meaning of the slur. The audience can
accept the presumably true de re proposition that that person likes Fox News. The audience can also accept that the speaker believes that poor white people are despicable; presumably the audience has learned this by hearing (1). Schlenker’s theory cannot explain the possibility that the non-bigoted audience still finds the slur strongly objectionable. In particular, it cannot explain why the audience is not now disposed to accept that poor white people are despicable for being what they are. It would seem that one of the explanatory advantages of a presuppositional account is lost – that audiences that accept and normalize the derogation of a certain group in discourse come to accept treating that group with contempt. Although Marques and García-Carpintero (2020) discuss this case, they do not address some of the other existing expressive-presuppositional views.

One could think that perhaps the problem with Schlenker’s theory is that it is an indexical view, referring only to the speaker of the context. Josep Macià (2014) has a different view. According to him, a use of “honky” presupposes that the speakers of the context (i.e., speaker and audience) are disposed to treat white people with a certain kind of contempt because they are poor and white. If the audience is not bigoted, this presupposition is false, and the falsity of the presupposition would explain why the audience is not willing to accept (1). We can grasp the point with a comparison with the presupposition associated with a pronoun, say, “she”. To accept an utterance with “she”, say (2) below, is to accept that the referent of the context is female:

(2) She’s not coming on the bike ride.

Of course, it may be false that the demonstrated person is female, or perhaps the demonstrated person is offended by being presented as female. The audience may thus resist accepting (2) if the presupposition associated with it is taken to be false. However, this solution still fails, as I will now argue.
To accept that the referent of the demonstrative pronoun is female (or accepts being presented as female) is to accept all that there is to accept in (2), beyond accepting the at-issue content that she is not going on a bike ride. But to accept the use of a slur is not merely a matter of accepting that we are disposed to treat the target of the slur with contempt. I can accept this as true for current purposes, and still not do what the use of the slur normatively requires – become disposed, or condoning, treating the targets of the slur with contempt.

The point may be made more forcefully. Suppose that the audience believes that they are all racists, because they have received diversity training at their workplace and have learned about structural and institutional racism. They now believe the proposition that, according to Macià, is presupposed by the use of the slur, and this is common knowledge. But the audience may nonetheless not condone treating a certain group with contempt and lack the disposition to react and behave that way (otherwise the diversity training would have been useless). So, the presupposition becomes common ground – it is treated as true. Presuppositions that become common ground need not be true, after all:

To accept a proposition is to treat it as true for some reason. One ignores, at least temporarily, and perhaps in a limited context, the possibility that it is false.
(Stalnaker, 2002, p. 716)

Thus, this account still does not capture what should be the main motivating feature of expressive-presuppositional views: to explain how people who accept the derogation of others in pejorative language come to accept the mistreatment of others. To accept the proposition that they are disposed to treat others with contempt is not sufficient for accepting treating certain people with contempt (the contrary is also true, but not relevant here).
1.2 Expressive presuppositions in richer contexts

Marques and García-Carpintero (2020) argue that discourse must update common ground by changing not merely the propositions that are commonly accepted, in the Stalnakerian belief-like sense, but by changing other aspects of contexts.

Several authors have argued that contexts must be structured illocutionarily. For instance, Roberts (2012, 2018) suggests that contexts are structured by a question under discussion (QUD), which discussants try to answer. Directives also help to establish the illocutionary structure of contexts, changing contexts by adding action plans (see Gibbard 1990, ch. 5, Charlow 2018, Portner 2016, and Garcia-Carpintero 2015, 2020).

As an illustration of how such a model might work, consider the account of expressive-normative discourse of Pérez Carballo and Santorio (2016)³. They modify Stalnaker’s model with an adaptation of Alan Gibbard’s (1990) norm-expressivism, where norm-acceptance is a practical attitude towards a possible course of action. A complete system of norms $n$, in contrast with a set of propositions $S$, determines a three-way partition of possible courses of action – those that are forbidden according to $n$, permitted but not required according to $n$, and those that are required according to $n$ (Perez Carballo and Santorio 2016: 611). Each assertion of a sentence is thus assigned a set of world-norm pairs. The two following examples that modify Schlenker’s example illustrate how sentences like those in (3) and (4) are assigned a set of world-norm pairs. A set of norms is only relevant to the content of the second sentence:

(3) Bob likes Fox News.

(4) Fox News is despicable.

(5) $\{\langle w, n \rangle: \text{Bob likes Fox News in } w\}$

(6) $\{\langle w, n \rangle: \text{Despising Fox News in } w \text{ is permitted by } n\}$
On this view, common ground includes both possibilities and norms that are left open by speakers’ attitudes: propositional doxastic acceptance states, and norm acceptance states, which would be action guiding.

But focusing on these kinds of attitudes is still insufficient for an account for pejoratives, and more generally for an account of hate speech, because neither pejoratives nor hate speech (need to) state which concrete actions plans are updating context. In fact, emotions can also be shared and motivate action, even before beliefs and intentions to act are formed. Rather, emotions can be prior and motivate beliefs and intentions, as I will indicate in the next section.

A currently prevailing view in the philosophy and psychology of emotion, these are normative appraisal states. Versions of this view about the nature of emotions have been argued by authors like Mulligan (1998, 2007), D’Arms and Jacobson (2000), Deonna and Teroni (2012, ch. 7, 2015). Emotions are intentional states. The emotion expressed in derogatory speech acts is contempt for the target of derogation, for instance. In saying that that one has contempt for a certain person or group is to ascribe to ascribe to the person or group a property, which is usually also called its “formal object”. The “formal objects” of emotions – say, being worthless or beneath consideration –, are the normative condition that appraise the emotional attitude given its content. If a person is beneath consideration, then to contempt would be the fitting attitude to have towards that person. If the person in question is worthy of respect as human being, then it is not fitting to have contempt. Now, formal objects are not part of the represented content, or what is represented, in this example, the person in question is the content of the attitude of contempt, but her appraisal as worthless or beneath consideration is not what is represented.
Now, if as authors like for instance Deonna and Teroni (2015) argue, emotions are attitudes with their own psychological profiles, and are evaluations that may be more or less fitting of their objects, then there are as many emotional attitudes as there are types of emotions (Deonna and Teroni 2015: 297). And if speech can be expressive of emotions, then different kinds of expressive discourse can express different attitudes, characterized by the specific evaluations that are proper of each emotional attitude.

This is the line that Marques and García-Carpintero follow. They argue that an expressive speech act is in turn defined by the expressed emotion’s related constitutive norms. If pejorative discourse is expressive of contempt, then “[i]n order to accommodate a presuppositional view of pejoratives, we should hence add further illocutionary structure to contexts. The intentional objects of emotional states are the contents of this additional structure.” (Marques and García-Carpintero 2020: 146) This is cashed out in terms of updates of context with the normative appraisals that identify the contempt expressed through derogation, which is presented thus:

*Derogation:* one must derogate group *G* on account of their having features *F*₁ … *F*ₙ only if group *G* has *F*₁ … *F*ₙ, and contempt fits a group with such features. (*Ibid.* 147)

To accept an utterance like (1), *that honky is watching Fox News*, is to accept the proposition that the demonstrated person is watching Fox News, but also to presuppose that the person is a member of a group with certain negative features, and to accept regarding members of that a group with contempt. An interlocutor who understands and accepts the utterance (1) thus accepts also that negative evaluation appraisal of groups presupposed to have certain negative features.

1.3 The pragmatics of harmful speech
This proposal goes some way towards filling in some of the gaps in existing accounts, including some that claim that speech can change attitudes that go beyond belief, such as Langton’s (2012). Langton (1993) had suggested that Austin’s (1962) distinction between an utterance’s causal effects and the act constituted by it can be deployed to argue that pornography can subordinate women, rather than merely having subordinating effects. This suggestion has been extremely fruitful for the development of work on harmful speech and hate speech. But Langton’s (2012) proposal is still programmatic:

I want to propose, in an exploratory spirit, the idea that the phenomenon of accommodation might extend beyond belief—beyond conversational score, and common ground, as originally conceived—to include accommodation of other attitudes, including desire and hatred. My remarks here will inevitably be programmatic. But to convey the general idea: just as a hearer’s belief can spring into being, after the speaker presupposes that belief, so too a hearer’s desire can spring into being, after the speaker presupposes the hearer’s desire; and so too a hearer’s hatred can spring into being, after the speaker presupposes that hatred. Stalnaker’s common ground can perhaps be extended to include not just common beliefs, and other belief-like attitudes, but common desires, and common feelings, as well. Speakers invite hearers not only to join in a shared belief world, but also a shared desire world, and a shared hate world. (Langton 2012: 86)

McGowan (2005, 2019) has further argued that language can oppress and constitute harm through its illocutionary force. Harmful speech will enact norms that prescribe harm and discrimination. Gelber (2019) follows Langton and McGowan, and says that

…these arguments support the premise that ‘hate speech’ can harm. In other words, hate speech according to these arguments can be understood as a discursive act of discrimination, which operates on its targets in constitutive and causal ways to effect the denial of equal opportunities and rights. (Gelber 2019: 6).

These authors agree that speech that is constitutively harmful should be distinguished from speech that has harmful effects. It is the former that should be labeled “hate speech”. But here I want to ask: how is the enactment of specific norms a constitutive harm of speech? The explanation that, for example, Langton gives of speech...
that is illocutionarily harmful does not establish that it is harmful because of its *illocutionary force* – but it is its illocutionary force that constitutes what it is as a type of speech act. Langton says, “according to the UN description, racial hate speech *disseminates* ideas that are based on racial superiority; it *promotes* racial hatred and discrimination. It also *incites*…” (Langton 2012: 75). The acts of disseminating, promoting or inciting are not constitutively harmful or hateful acts. So it is not what an act is constitutively that characterizes it as harmful, at least not if we’re talking of these acts. It is thus understandable that the alternative explanation is to say that it is the *contents* that are promoted, incited, or disseminated that are harmful or discriminatory. However, we need to say more about why these contents are harmful, or wrongs.4

McGowan argues that it is through the norms that are enacted that speech can be harmful and discriminatory. But I think that it also falls short of establishing that this harm is one that characterizes speech as *constitutively* harmful. McGowan says that speech constitutes harm by enacting norms that prescribe that harm (2019: 20). Constituting harm is thus a specific way of *causing* harm. She offers two examples of discrimination against red-haired people to illustrate this difference. One case involves causing discrimination against red-haired people, and the other is meant to illustrate harm caused by the enactment of a hiring policy:

Suppose instead that I am an employer and implement a company hiring policy when I say, “From now on, we no longer hire anyone with red hair”. This utterance will cause discriminatory conduct on the part of my employees but it does so via the prescriptive force of the hiring policy enacted by my utterance. (McGowan 2019: 24)

There are details about this story that are not laid out, although we are invited to interpret it in a way that makes it *true* that the policy against hiring red-haired people is unfair differential treatment. But it *need not be*. And this makes it a bad explanation of
constitutively harmful speech. Suppose the job in question is for lifeguards on a beach on a country on the equator, say, São Tomé and Príncipe, where people prone to develop skin cancer should avoid sun exposure as much as possible. The employer’s decision to not hire people prone to get skin cancer (which includes red-haired people) in that line of work is not unfair, although it is differential treatment. So there is nothing constitutively harmful in the speech act in question – it is an act that enacts a norm – and there is nothing about the content of the norm as such that is harmful, hateful, or discriminatory. Whatever features there are of enacting a norm with that content in another context that makes it a discriminatory act are features that will be explained by things extrinsic to speech, its force, or its content. I think this means that these pragmatic accounts fail to explain why the harm caused is constitutive of what the speech act performed.

Now, if Marques and García-Carpintero are right in claiming that the norms enacted norms through expressive discourse are provided by the kind of emotion that is expressed, then this may count as a good basis for an account of speech that is constitutively harmful. For instance, derogatory speech is constitutively harmful by enacting norms like Derogation.

In the next section, I will briefly draw from work on collective emotions in joint action and in politics, and about hate more specifically, to suggest that it can be combined with this framework of expressive discourse and give a better conceptual understanding of hate speech.

2. Negative emotions and the hate sentiment

In this section, I (plan to) combine an account of the illocutionary structure of conversational contexts and of the normative requirements that speech makes on context
(Roberts 2012), and recent accounts of the attitudes or sentiments expressed through hate speech (Fischer et al 2018, Halperin et al 2012, Salmela and v. Scheve 2017).

The kinds of acceptance attitudes that are typically assumed to play a part in conversations and in collective actions are beliefs and intentions. However, as the previous section argued, in order to explain derogatory or hate speech we need to take into account other attitudes and dispositions. To appreciate, we should understand the need to distinguish between collective intentions and collective emotions.

The idea that plans can be shared and be common ground is a familiar one in the philosophy of action and in social ontology. For example, on Michael Bratman’s theory, collective action requires individuals to share their intention to act together in a shared cooperative activity (Bratman 1992, 1999, 2009). But Bratman’s theory has received criticism. It requires shared plans to be common knowledge among the individual agents who act together. This is however a very strict and over intellectualized constraint to place on any possible collective, not to mention large or unstructured collectives. Individuals can also do their part of collective activities without intending to bring about the aims of the joint action (Kutz 2000). Employers or workers in large corporations, followers of political movements, or citizens of a nation or region, don’t meet Bratman’s conditions for collective agency.

There are alternatives to Bratmanian theories. Salmela and Nagatsu (2016) offer an account of collective emotions that links the intentional structure of joint actions and the underlying cognitive and affective mechanisms. They show that emotions can function as both motivating and justifying reasons for jointly intentional actions. Crucially, emotions can precede intentions to act, and can motivate the formation of intentions and of beliefs.

Emotions can be shared and spread among a community:
[w]e suggest that these [collective] emotions emerge when a group of individuals appraise the emotion eliciting event convergently in relation to their overlapping private concerns, and such mechanisms as attentional deployment, emotional contagion, facial and motor mimicry, and behavioral entrainment synchronize the individuals’ emotional responses, producing a shared affective experience among individuals who are mutually aware that others are feeling the same.” (2016: 8)

García Rodríguez (2021) has recently argued that emotions can be perceived in other people’s emotions in their bodily and behavioural expression. So, although emotions and their formal objects are not conveyed in the way that the contents of beliefs are, that does not preclude our capacity to perceive and to share affective experiences with others.

There are various perspectives about what hate is. Matsumoto et al (2016) say that recent theoretical and empirical work suggests that the combination of anger, contempt, and disgust (ANCODI) are the basic elements of hatred, and that these are the key emotions associated with intergroup aggression. Fischer et al. (2018) conducted an extensive review of existing research on hate. As they say, hate is developed when others mistreat or humiliate someone, or whose deliberate actions have become an obstruction to someone’s goals:

…when individuals experience hate, they typically perceive their hate target as having malicious intentions and being immoral, which is accompanied by feelings of lack of control or powerlessness. Such appraisals are not the result of one specific action, but of a belief about the stable disposition of the hated person or group.” (Fischer et al. 2018, p.310)

They acknowledge that hate is hard characterize. It can be an episodic emotion, directed towards someone who is perceived to be malicious. More importantly for my present concerns, hate can also evolve into a long-term sentiment, particularly in intergroup contexts. Someone who feels hate as an episodic emotion against another person can be hostile or threaten violence against the person, but if it’s an episodic emotion against an individual, and the other person’s social standing or safety is not
threatened, an expression of that emotion, say a child shouting to his friends who don’t want to go on playing “I hate you!”, is not hate speech.

The child shouting “I hate you!” to his friend expresses the child’s emotion. Yet, what an adult minder who hears what the child says is that the child is angry or upset at with the friend. Her expression of the episodic hate towards other children does not produce in the minder the same hate emotion that the child feels. This is not what happens with hate speech. Its danger is that an audience sympathetic to the speaker can come to accept or condone hatred towards the target in that context.

It is not episodic hate emotions, but long-lasting hate sentiments, that create the systematic conditions for hate speech to target a group. These systematic speech effects occur in a wide range of contexts and are hard to overcome or escape when they occur. Halperin et al. (2012) say that the hate sentiment in intergroup contexts is a stable and familiar “hating” emotional attitude (“chronic hatred”), which organizes people’s social world and helps strengthening the connection to the in-group (“in-group love”) at the expense of various out-groups (“out-group hate”). The goal of the hate sentiment is to eliminate the out-group from the in-group’s environment, for example through an absolute separation from members of the other group.

Some recent research shows that hate spreads from the individual to the collective level and is shared by ingroup members (Jasini and Fischer 2018). Moreover, the increase in hate speech against, say, immigrants, correlates with the increase in hate crimes against them (Salmela and von Scheve 2017).

If hate speech is expressive of hate sentiments, then it can explain some of the systematic effects of hate speech, a feature that Gelber (2019) wants to account for. As Maitra says (2009), systematic speech effects occur in a wide range of contexts and are hard to overcome or escape. Such speech effects are presumably conventional aspects of
meaning, and one explanation for this is that they are presuppositional, as claimed here. Hate speech, then, is expressive of hate sentiments against specific target groups, and the norms that constitute those sentiments can update what is common ground, i.e., what is taken for granted, in society or among an in-group.

Other alternatives would not work. We can’t represent hate sentiments that become common ground merely in terms of the acceptance of a presupposition that a speaker hates persons G. This does not explain the spread of hate through hate speech. So, while assertives add propositions that are taken for granted as common ground, and directives and interrogatives add plans and QUDs, expressives add contents that support emotional and reactive attitudes shared by the interlocutors. In felicitous cases, these different commitments are mutually shared and license presuppositions, which motivate shared actions.

The normative requirements that hate speech makes on common ground are provided by the characteristics of hate sentiments, namely:

(i) negative appraisals of outgroup members as malevolent or malicious just by being members of that group,

(ii) action tendencies that go from revenge, social exclusion, or attacks, to the destruction of the target group, and

(iii) motivation goals, such as the desire to harm, humiliate, or even kill the target (from Fischer et al 2018, 311).

When hate speech occurs and is tolerated, it updates context with the contents that make hate ‘fitting’ or ‘appropriate’ (Deonna & Teroni 2012, ch. 7). The characteristic action tendencies and motivational goals generated by hate sentiments go to another illocutionarily distinct part of the context (QUDs and plans). And if accommodated, these appraisals are taken for granted and shared by interlocutors.
3 Objections, replies, and some conclusions

The framework offered here resists some of the existing objections against hate speech being expressive of hatred. I focus on the systematic criticism to the claim that hate speech expresses hate. In two papers, Alexander Brown (2017a, 2017b) argues that it’s a myth that emotions, feelings or attitudes of hate or hatred are part of the essential nature of hate speech (2017a). He also argues for a conceptual analysis of the ordinary concept *hate speech* as a family resemblance concept (2017b). I will focus on the negative arguments rather than his positive proposal. Brown says,

> Focusing on the assumption that emotions, feelings, or attitudes of hate or hatred are one basic building block of the ordinary concept hate speech, I proceed to argue that despite the existence of several possible connections between speech and hate or hatred none appear to capture an essential feature or defining quality of hate speech. (Brown 2017a: 433)

Brown argues against an analysis of the concept *hate speech* in terms of these three conditions the speech itself or other expressive conduct, groups or classes of persons identified by protected characteristics, and emotions, feelings, or attitudes of hate or hatred. He then disambiguates four readings of ‘hate speech’ or ‘expressive of hate’ and argues against each.

Before proceeding, I must make a preliminary remark about the focus on groups identified by protected characteristics at the level of a general definition. I do not think that hate speech should be primarily understood as speech that targets protected characteristics. There are historical cases where hate speech correlates with hate crimes and mass violence. The similarities of the cases transcend local systems of discrimination or oppression, different histories of unresolved conflicts, and cases where the targets of violence and hate speech are groups that were not vulnerable or subordinated. Some
analyses of harmful speech focus on pre-existing discrimination (for instance, Kinderman (MS)).

Jonathan Maynard and Susan Benesch (2016) offer an account that describes dangerous speech as speech that is capable of encouraging approval of violence by an audience: “The forms of speech and ideology that catalyze mass violence, and the ways in which they do so, are strikingly similar across different cases” (Maynard and Benesch 2016: 71). Dangerous speech, as they characterize it, is inflammatory speech that is motivational. Now, Maynard and Benesch don’t seem to give a definition of “inflammatory speech”. They avoid using “hate speech” given the plurality of legal and theoretical definitions. But I will assume that by “inflammatory speech” they mean what I want to characterize as hate speech. This means that we can understand their project as a study of the conditions in which hate speech correlates with hate crimes, in general.

Some of the cases they address include the genocide of the Tutsis in Rwanda, or the violence against intellectuals, political dissidents, or kulaks (farmers that owned property) of the Stalinist era in the Soviet Union. None of these are adequately described as groups identified by “protected categories”, prior to the spread of hate speech against them. (nonetheless, the requirement of “protected categories” may be necessary to draft actual policies to protect groups that we know are vulnerable).

What I want to emphasize is that if we include in our analysis the reference to characteristics that are already recognized as protected — then we will be unable to use our theory to tackle and predict cases where speech correlates with episodes of hate crimes against groups that were not previously identified by “protected characteristics”. Since these cases exist and involve some of the worst crimes against humanity, any theory of hate speech that does not explain its contribution to hate crimes has fatally failed in its
explanatory aim. I should emphasize that it is only in specific circumstances, as Maynard and Benesch illustrate, that speech becomes dangerous.

Let us return then to Brown’s objections to the essential connection between hate and hate speech. He correctly points out that the connection between emotions and discourse can be read in different ways. Under one understanding, hate speech is speech that is perceived as repugnant or as hateful by an audience:

S is hate speech iff it is:
(1) a form of speech,
(2) about or against members of groups or classes of persons identified by protected characteristics, which
(3a) arouses reflexive hatred amongst some significant section of society.

To spell out the last clause, Brown says “the putative connection between speech and hate consists in the fact that a significant section of society hates certain forms of speech that are about or against groups or classes of persons identified by protected characteristics.” (Brown 2017a: 447)

The problem with this analysis is that hearer reactions are not reliable guides to assess hate speech, and audience’s reactions are not a good basis on which to regulate speech. But fundamentally different people have varying sensibilities, and regulating speech on the basis of individual sensibilities is inconsistent with the free speech doctrine.

The second possible reading of the third clause, according to Brown, understands it as speech that signifies, articulates or otherwise represents the speaker’s emotions. This can be called ‘hate-filled speech’ or ‘hate-laden speech’:

S is hate speech iff it is:
(1) a form of speech,
(2) about or against members of groups or classes of persons identified by protected characteristics, which
(3b) carries forth, articulates, symbolizes or represents in some way the speaker’s emotions, feelings, or attitudes of hate or hatred toward the subject of the speech.
As the reader can gather, the claim I advocate is close to the second reading, and so I’ll return to this later.

The third possible reading understands it as speech that is motivated by the hatred felt by the speaker “for any members of groups or classes of persons identified by protected characteristics”. This is a notoriously poor interpretation of hate speech.

S is speech iff it is:
(1) a form of speech,
(2) about or against members of groups or classes of persons identified by protected characteristics, which
(3c) is speech motivated by hatred of the speaker. (Brown 2017a: 455-6)

This is a bad alternative because it overgenerates and counts as hate speech that is in no way harmful, discriminatory, insulting, threatening, etc. For example, a person who is embarrassed and ashamed of his hatred of strong independent women goes out of his way to monitor what he says, and only says positive things about them. It is reasonable to regard this as speech that is motivated by the speaker’s hatred, but it is not hate speech.

On the final reading, hate speech is speech that incites hatred toward any members of groups or classes of persons identified by protected characteristics.

S would be hate speech iff it is:
(1) a form of speech,
(2) about or against members of groups or classes of persons identified by protected characteristics, which
(3d) intentionally, or has the likely outcome, of inciting hatred. (Brown 2007a: 458)

Depending on how we understand this fourth reading of the third clause we may either find straightforward objections, as I’ll now consider, or find a view compatible with the expression of hate.
The objections to this interpretation, as it stands, are fairly straightforward. On the one hand, thoughtful reasonable speech can have as an outcome an increase in hatred. For example, the announcement of an electoral result where supporters of a party A lose to another party B, after a long and difficult polarized campaign, may incite hatred in members of A against supporters of B, but it should not count as hate speech. On the other hand, and perhaps more controversially, not all speech that is said intentionally with the aim of inciting hatred should count as hate speech. For example, a detailed description of the crimes committed by a group of individuals against an innocent person, offered to irascible members of the family of the victim with the intention of inciting hatred against the criminal gang, is still not hate speech, even if it is said with the intention of inciting hate. On the other hand, Brown argues that using a derogatory word in private to a black person without any intention to derogate, and without inciting hatred against black people in general, can nonetheless be hate speech.

Nonetheless, this final reading in terms of incitement has some weight. The UN International Convention for the Elimination of all Forms of Racial Discrimination, article 4, condemns all attempts to justify or promote racial discrimination (see also Langton 2012: 74-5). The International Covenant on Civil and Political Rights, article 20(2) says “Any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence shall be prohibited by law.” I think that there is a way of reading these statements consistently with the expression of hate.

Let me then reply to Brown’s objection that hate speech expresses hate. He rejects the thesis that hate speech involves the expression of the “emotions, feelings, or attitudes of hate or hatred of the speaker for any members of groups or classes of persons identified by protected characteristics”.


When this sort of charge is brought against expressivist views, it seems to assume that the description “the emotions, feelings, … of the speaker” is to be read as a *de re* definite description, such that, if the speaker does not feel hate, say, then the description would fail to pick any set of attitudes. Presumably, in that case, since the speaker has no hate, then he is not expressing hate, and therefore his speech should not count as hate speech. Given that the examples appear to be cases of hate speech, it should follow that expressivism is false. Now, this argument is no more problematic for the expressive-presuppositional account that I defend here than the analogous argument to the effect that derogatory discourse cannot express contempt. The first section of the present paper showed that this is a mistaken conception about expressive speech in general, and the second section supported the idea that hate emotions and hate sentiments can differ in sufficient ways to allow for one to have a dispositional hate sentiment even if not feeling the phenomenology of episodic hate emotions. That is not to say that there aren’t hard cases that escape an easy classification (see Mikkola (MS) on hard cases that *should not* count as hate speech). Brown gives essentially three types of examples, which I now present and discuss.

*First case: Fundamentalist Christian.*

A fundamentalist or evangelical Christian directs the following words at people on a public street as they enter an LGBT social club, “You homosexuals and lesbians are sinners in the eyes of God, you disobey the teachings of the Bible, and for this reason you will go to hell if you do not repent.” Brown says that this utterance, conceivably, is not “an articulation or expression of emotions, feelings, or attitudes of hate or hatred” but rather “an expression of deeply held religious beliefs or as some sort of religious exercise or rite of passage or even as an outpouring of feelings or sentiments of disgust or repulsion that have been learned from parents or community leaders.”
There are two kinds of issues here. First, suppose that this is an expression of religious belief. In so far as the utterance relates to homophobic hate speech, it does so because it expresses the presumed negative features that the parents or community leaders suppose are essential features of LGBT people, and that would make the hate of LGBT people an apt or fit emotion (in the eyes of said evangelical or fundamentalist Christian people). Similar things may be at stake (depending on the speaker, audience, and other features of context) in some of the cases mentioned earlier, where some negative stereotypically claims are made about a particular group. Therefore, although stating that a group has certain stable and essential negative features is not equivalent to expressing hate for the group, such claims can come to expressively presuppose hatred for the group, for instance in contexts where those statements have become conventional means of giving reasons for despising members of the group and where expressions of hatred against the group are also widespread.

Second, the “outpouring of feelings or sentiments of disgust or repulsion”, if accompanied by further emotions, such as contempt, anger, or fear, may well indicate not only that hate is expressively presupposed, but that, furthermore, the person actually hates LGBT people. Brown additional remarks about the example that “speech that carries the prejudiced (because contemptuous) message that homosexuals are morally inferior beings or as a speech act that amounts to ranking, degrading, harassing or persecuting homosexuals, it might qualify as hate speech” (Brown 2017a: 452). But these remarks are consistent with the expression of homophobic hate, because they are compatible with the negative appraisals and motivations that distinguish hate from other sentiments.

Now, if the normative requirements of expressive hate speech are provided by the characteristics of hate sentiments, then not only are LGBT people negatively appraised as malevolent or malicious -- “You are sinners and you’ll go to hell!”, but certain action
tendencies become permissible (or required) in light of the negative appraisal, and motivational goals becomes desirable – action tendencies liek revenge, social exclusion, attacks, destruction, and motivational goals like the desire to harm, humiliate, or even kill the target (from Fischer et al 2018, 311).

Second case: Scientist

Imagine a scientist who publishes a controversial paper with the results of his research about the relative intelligence of African Americans, and which includes the sentence, “African Americans tend to have lower IQs than white Americans.” As Brown says, it is

…perfectly intelligible for people, not only African Americans, to think that this is hate speech by virtue of the fact that in a very public way it rehearses and supports, rather than challenges and confronts, a particular negative stereotype or social stigma about African Americans as unintelligent or less intelligent than whites. (Brown 2017a: 452)

This is case falls under what Mikkola calls “hard cases”, which include similar cases like a scholarly paper arguing that the Holocaust never happened. I agree with Mikkola that, in general, hard cases ought to be classified as different types of prejudicial speech. She follows Brink (2001) in arguing that cases of this kind may count as discriminatory speech, in that “they reflect group stereotypes and represent groups or their members as inferior by virtue of these stereotypes.” (Brink 2001: 133). However, it is possible that these kinds of prejudicial speech occur in contexts where hate speech about those groups is also common, and, in which case, a claim negating that the Holocaust occurred, uttered for example by a neo-Nazi leader to his followers, may be considered as hate speech.
It is possible that some disablists could use the insults ‘retard’, ‘cripple’, and ‘freak’ in order to express their contempt or disdain for people with physical or mental disabilities, to express the fact that they look down on such people, believing them to be worthless or beneath consideration, without at the same time articulating, symbolising or representing emotions, feelings, or attitudes of hate or hatred toward such people. But, like previous cases, the expression of contempt and disdain, together with the belief that the people in question are worthless or beneath consideration, would count as expressions of hate in my view.

References


____ (2014). Expressive Meaning and Presupposition. Handout for a talk at the Names, Demonstratives, and Expressives conference in Gargnano, Italy.


Mikkola, M. (MS). Discriminatory vs hate speech: wherein lies the difference? (manuscript).


Boisvert’s expressive-assertivism seems to be preferable to other hybrid theories, like Barker’s (2002) or Ridge’s (2006). This discussion is beyond the scope of this paper.

Boisvert relies on the distinction between direct and indirect illocutionary acts when he argues for expressive-assertivism. Indirect illocutionary acts are those acts that are performed indirectly, by means of the performance of another act. For instance, by asking the question “Could you pass the salt?” the speaker can request that the interlocutor pass the salt, i.e., by performing an interrogative act the speaker can indirectly perform a directive. Direct speech acts, then, are those that are not indirect.

For a discussion and criticism of some aspects of Pérez Carballo and Santorio’s proposal, see also Marques (2021).

Similar criticism of Langton’s proposal is offered by Daniel Jacobson (1995) about silencing as illocutionarily disablement. Jacobson argues that the disablement of some illocutionary acts need not be a wrong, even if it is systematic.