Understanding and appraising 'hate speech'

Sara Vilar-Lluch, University of Reading | University of Nottingham

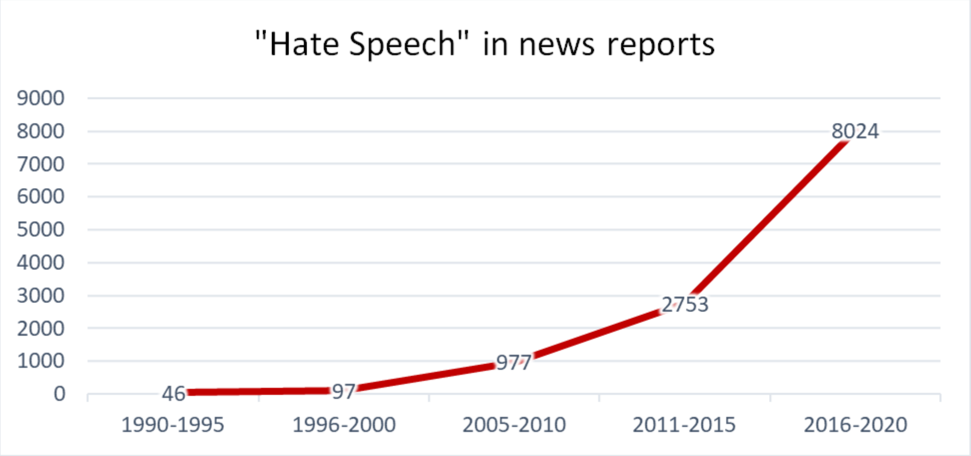
Hate speech has become a matter of international concern, permeating institutional and lay discussions alike. Yet exactly what it means to refer to a linguistic act as 'hate speech' remains unclear. This paper examines the lay understanding of hate speech, focusing on (1) the relationship between hate speech and hate, and (2) the relationship between hate speech and offensive speech. As part of the second question, the paper considers how hate speech is defined as a legal matter in the UK Public Order Act 1986. The study adopts a corpus-based discourse analysis approach and examines 255 hate speech-related news articles and the general *English Web 2020* corpus. Hate speech is a complex multifaceted phenomenon; while 'hate' is one of its core characteristics it is not sufficient to assess a certain behaviour as hate speech. Threats, denigration of the targets based on a protected characteristic (age, race, religion, sex, sexual orientation, disability), the potential to cause harm and the intent to stir up hatred are also essential in distinguishing hate speech and offense.
Keywords: hate speech, ordinary meaning, legal discourse, offensive speech, corpus linguistics

1. Introduction

The term 'hate speech' first emerged in the 1980s among legal theorists who sought to deploy legal measures to counter harmful racist utterances (Delgado, 1982; Matsuda, 1989). Failing to regulate hate speech ("words that wound") perpetuates discrimination through inaction, implicitly attributing a different value to different individuals (1982, 141; 1989, 2322-2323). Since then, 'hate speech laws' have been adopted across a number of jurisdictions, although not all are labelled as such. For instance, the UK's Public Order Act of 1986 prohibits the use of "threatening," "abusive," or "insulting" words when these words are either intended to "stir up racial hatred" or are likely to do so. These legal descriptors emphasise the illocutionary character of hate speech: a speech act is not 'hate speech' for communicating hostility, but for "act[ing] *upon* its addressee in an injurious way" (Butler [1997] 2021, 16).

Over the last two decades, the term 'hate speech' has moved beyond the legal realm, forming part of ordinary discourse about speech that vilifies and incites hatred or violence against vulnerable groups (Brown, 2017a, 422-424). As an illustration, Figure 1 shows the evolution of the news coverage of the topic in British newspapers, as retrieved from Lexis Library (Nexis database). Figure 1 reveals an exponential rise starting in the first decade of the 2000s, which can be partly attributed to the popularisation of social media platforms, with social media policies, incidents, and trolling forming recurrent topics across the news reports of the last ten years.

Figure 1 Hate speech-related news reports in British newspapers as retrieved from Lexis Library (Nexis)



Responding to growing social alarm about hate speech, a number of linguistics and computational linguistics studies have emerged to detect and tackle the use of hate speech in social media sites (e.g., Basile et al. 2019; Sanguinetti et al. 2018). However, any operationalisation of hate speech identification requires establishing and filtering hate speech-related words, and this involves decision-making regarding what should count as hate speech. These decisions are not straightforward. Hate speech is a complex phenomenon and without a better understanding of how ordinary speakers appraise it, there is a risk that the instances identified as hate speech by automated hate speech detection systems will not match ordinary speakers' judgements in this area. This would be problematic, not simply because it would undermine public confidence in AI-solutions, but also because of the potentially chilling effect of too inclusive a definition of hate speech. If communicative acts that the majority of people would not count as hate speech are classified as hate speech by the detection systems, warranted expressions of complaint (perhaps from oppressed groups directed at more powerful groups) or 'insurrectionary acts' (Butler [1997] 2021, 145) may be silenced due to the speaker's concerns about being identified as a perpetrator of hate speech. However, broad definitions of 'hate speech' have been suggested. For instance, Schmidt and Wiegand (2017, 1) argue for an all-embracing understanding of the concept, defining it as "a broad umbrella term for numerous kinds of insulting user-created content". This leads them to categorise the following examples as instances of web data coded as hate speech (Schmidt and Wiegand 2017, 1):

> (a) Go fucking kill yourself and die already useless ugly pile of shit scumbag.
> (b) Hope one of those bitches falls over and breaks her leg

That (a) and (b) illustrate highly offensive speech directed towards a target is beyond doubt. But do they constitute instances of hate speech? To answer that (and related) questions, we need to explore how ordinary speakers understand 'hate speech' and ascertain the degree to which this ordinary understanding matches (or fails to match) the legal understanding of the term.

The next section (2) considers the relevance of examining the ordinary meaning of ‘hate speech’. The paper then offers an overview of the data and methods (section 3), followed by two analytical sections: section 4 examines the relationship of ‘hate’ and ‘hate speech’ and section 5 the legal understanding of the concept. The paper closes with some final considerations on the ordinary and legal understandings of ‘hate speech’ vis-à-vis offensive speech.

2. A case for studying the ordinary meaning of ‘hate speech’

A better understanding of ‘hate speech’ is important for practical and principled reasons: it may help to address the difficulties that arise from a misalignment between legal and other restrictions, and the ordinary meaning of hate speech; and it can shed light into the principles that govern legal interpretation and explore any risk of violation of fair notice. These points are discussed in turn in the following sections.

2.1 Practical reasons for studying the ordinary meaning of ‘hate speech’

A primary driver for wanting a better understanding of ‘hate speech’ is to inform and monitor the development of hate speech detection software. If these systems operate in ways that fail to align with the ordinary understanding of ‘hate speech’ (either failing to identify communicative acts ordinary speakers would categorise as instances of hate speech, or identifying as hate speech communicative acts the average speaker would not treat as such), they will fail to perform the role they were designed for, undermining public confidence and potentially producing an unwelcome chilling effect on ordinary speech. Potential misalignment with ordinary usage seems to be more than a theoretical risk, since a range of different definitions is adopted by scholars in this area. Warner and Hirschberg (2012, 19) understand hate speech as a “form of offensive language that makes use of stereotypes to express an ideology of hate” and, following their definition, establish that unnecessary mentions of race or ethnicity and labelling of individuals as belonging to a group should be categorised as hate speech (2012, 20). In contrast, Sanguinetti and colleagues (2018, 2799) opt not to include stereotyping alone as a sufficient condition for hate speech detection. Instead, they distinguish “offensive speech” from “hate speech”, and propose an annotation framework that includes stereotyping, aggressiveness, offensiveness, irony and hate speech. Hate speech is defined in terms of the illocutionary force of the utterance and its target, the latter described as a member of a group identified as recurrent hate speech victims (2018, 2800). Projects establishing hate speech filters frequently discuss having had to make decisions on a range of potentially confounding features, such as cultural variation, individual biases, and the distinction of hate speech from offensive language, threats or bullying, to mention but a few (Sanguinetti et al. 2018, 2800; Warner and Hirschberg 2012, 20).

2.2 Theoretical reasons for studying the ordinary meaning of ‘hate speech’

Several cannons of legal interpretation make reference to the notion of ‘ordinary meaning’, particularly in countries governed by English Common Law. Following

English Common Law, legal interpretation is guided by the Plain Meaning Rule, which establishes that in cases of unambiguity, statutory texts should be interpreted according to the ordinary meaning of language, excluding any possibility to refer to 'extrinsic' evidence in interpretation except when this would lead to absurdity or injustice (Manning 2003, 2388-2389; Mouritsen 2011, 159-160). Traditionally guided by dictionary searches, some legal experts and scholars have recently turned to corpus approaches to illuminate the ordinary meaning of legal terms and inform judges' interpretations (see Lee and Mouritsen 2018, 2020; Mouritsen 2011; Solan and Gales 2016). An underlying motivation for considerations of ordinary meaning is the principle of fair notice: individuals should be able to understand the laws that govern them. In the context of 'hate speech', this requirement for fair notice brings to question the extent to which the ordinary understanding of 'hate speech' matches the legal understanding. It is important for fair notice that the acts categorised as hate speech within the law are broadly identical with the instances categorised as hate speech in ordinary parlance.

According to many hate speech laws, a key element of 'hate speech' is the promotion of *hatred* (Brown 2015, 26-28). In the UK Public Order Act 1986, hate speech is referred to as speech that is liable to stir up 'racial' or 'religious' hatred, or 'hatred on the grounds of sexual orientation'. Commenting on international jurisprudence and the Canadian law, Rikhof explains that 'hatred' "connotes emotion of an intense and extreme nature that is clearly associated with vilification and detestation", and that "only the most intense forms of dislike fall within the ambit of this offence" (Rikhof 2005, 1126, 1131). The association of hatred with the nature of 'hate speech' is common among legal scholars (see Brown 2017a, 431, 436 for a discussion) and, to a point, current linguistics research seems to support this identification (see Culpeper 2021, although in Culpeper's study the focus is on 'hateful', not 'hatred'). While, 'hateful' can mean being 'filled with hatred', 'full of hate' or 'arousing hate' (see *Merriam-Webster*), it seems that Culpeper's investigation assumes that there is a straightforward connection between hate speech and hatefulness. Understanding the connection with hate as decisive for hate speech, would align with Schmidt and Wiegand's (2017) decision to code examples (a-b) above as instances of hate speech, since the speakers were explicitly displaying hatred towards the addressees. However, in his seminal article about the ordinary meaning of 'hate speech', Brown (2017a, 439) argues against the ordinary association of 'hate speech' with speech that is connected to feelings of hatred, understood as "extreme dislike or aversion" toward the victim. One question the present study needs to address is the degree to which the ordinary understanding of 'hate speech' relies on the feeling of hate for counting as 'hate speech'.

Finally, the theoretical and practical significance of studying the ordinary meaning of 'hate speech' goes beyond our understanding of hate speech alone. In linguistics, there has been a growing interest in studying (im)politeness concepts as first-order terms, i.e., as they are commonly understood by the social actors (e.g., Culpeper 2021; Culpeper and Haugh 2021; Taylor 2015, 2017). Understanding how speakers of a language conceive concepts such as "offence", "hate speech" or "rude" is of paramount importance for studies of linguistic behaviours. Studies on language use recorded as hate crime (e.g., Culpeper et al., 2017) make it possible to further examine whether court decisions align with public perceptions of hate speech. This study can contribute to the growing area of research on the metalinguistics of impoliteness-related concepts. While recognising that providing a full analysis of the meaning of 'hate speech’ is beyond the scope of any single paper, in what follows this paper shows how the use of corpus methods may shed light on the following questions: (1) how is ‘hate speech’ related to hate? (section 4) and (2) how is ‘hate speech’ related to offence? (section 5).

3. Data and methods

This study combines corpus linguistics with qualitative discursive analysis, the latter focusing on metaphor use and the expression of evaluation to better account for the relation of hate speech with the emotion of hate. Corpus methods are the preferred approach for metalinguistic studies on the lay understanding of terms (see Culpeper 2021; Culpeper and Haugh 2021; Taylor 2015, 2017). A corpus approach offers empirical validity and allows access to naturally occurring data. Certainly, corpus studies are non-exhaustive of the meaning of the terms examined: the reported uses in the corpus reveal part of the meaning that a linguistic community attributes to the term, but do not exhaust meaning (Culpeper 2021, 7; Haugh 2016, 45). To face this challenge, metalinguistic studies usually build large corpora (Taylor 2015, 2017) or use a general corpus (Culpeper 2021; Culpeper and Haugh 2021).

However, studying the ordinary meaning of ‘hate speech’ comes with a methodological concern: while language users use words such as “rude” or “irony” to describe their speech, ‘hate speech’ is mainly a second order term (Culpeper 2021, 8); speakers ‘mention’ the concept, usually reflecting legal uses and legislation purposes, but do not normally use it to characterise their speech acts. In order to circumvent this issue, a small corpus of news articles about hate speech was built for the purpose of the study (‘journalistic corpus’). Results of the pilot analysis of the journalistic corpus guided searches into the general corpus *English Web 2020 (enTenTen20)* available in *Sketch Engine* (Kilgarriff et al. 2014), the software employed for the analysis.

The journalistic corpus includes 255 news articles comprising a total of 164,183 words. The articles date from 1990 to 2021 and were published by British national newspapers (e.g., *The Guardian*, *The Times*, *Daily Mail*) and regional ones (e.g., *Yorkshire Post*, *Belfast Telegraph, Birmingham Post*). The articles were retrieved from the Lexis Library News of Nexis database and had been tagged as being, at least, 70% hate speech related.

Journalists, especially those specialised in legal matters, may be more versed with hate speech laws than the general public, thus potentially differing from lay people’s understanding of ‘hate speech’. Divergencies notwithstanding, the journalistic corpus is relevant for two main reasons. The different communicative goals and audiences of news articles and legal registers require different linguistic choices, and uses of the terms may not coincide. Register differences can also condition public reception; research has suggested that when legal terms are used in non-legal contexts readers are less inclined to attribute them a legal meaning (Tobia et al., 2023, 27). On the other hand, news media has the capacity to inform public judgements of the reported phenomena (e.g., Fairclough 1989; Fowler 1991), hence influencing lay people’s understanding of hate speech.

The *enTenTen20* corpus comprises 36 billion words of texts collected from Internet domains of countries with English as official language; texts cover a variety of topics (arts, business, games, health, home, recreation, reference, science, sports, society, and technology) and a variety of genres (blogs, discussions, news and legal).[1] A corpus with a higher presence of social media discourse would help provide a more representative picture of the social perception of hate speech. However, the variety of text types of the *enTenTen20* already makes it possible to infer whether the results of the pilot reflect journalistic idiosyncrasy or can be extrapolated to the general English-speaking community.

The analysis of the journalistic corpus involved a study of the keywords, followed by an analysis of the collocates of "hate speech", "hate", "hatred", "incitement" with the Word Sketch tool of *Sketch Engine*, which organises the results according to the grammatical patterns associated with the node (the term analysed), hence reflecting how the collocates operate in context. Collocation strength is expressed throughout this paper by logDice score, an association score that works well with different corpus sizes (Rychlý 2008). The collocates analysis was complemented by examining corresponding concordances. Following the pilot analysis, collocates for "hate speech", "hatred" and "incitement" were examined on the *enTenTen20*. The pilot results guided further corpus searches considering hate speech behaviours, countering actions, and the association of hate speech with hatred. Collocates analysis was refined with a qualitative examination of concordances, which considered the expression of evaluation and metaphor use.

The study adopts a discursive approach towards metaphors (e.g., Semino 2008; Musolff 2016). The analysis of evaluation is based on Martin and White's (2005) Appraisal framework, which distinguishes between the following attitude types: Affect (expression of feelings), Judgement (evaluation of human behaviours) and Appreciation (evaluation of 'things', performances and natural phenomena). These attitude types are further subdivided in a more delicate classification and are distinguished between inscribed (explicit) and invoked (indirect), according to the degree of explicitness of the evaluation (Martin and White 2005, 45-58). For simplification purposes, this study does not distinguish level of inscription.

The study of the ordinary understanding of hate speech is complemented with an analysis of the semantics of the terms "threatening", "abusive" and "insulting", used to describe hate speech behaviour in legal-related documents. Collocates are examined focusing on the usual targets of actions qualified as "threatening", "abusive" and "insulting", entities associated with such descriptors, and characteristics related to them. This analysis can show us whether the legal description of hate speech accounts for the ordinary understanding of the phenomenon and indicate possible enhancements.

4. Hate and hate speech

4.1 When hate is not an emotion

[1] Legal is the least represented genre, which contrasts with discussions and news. Information about the *enTenTen20* general corpus is available at: https://www.sketchengine.eu/ententen-english-corpus/

Responding to the characterisation of hate speech as being hateful and promoting hatred, we examined how the emotion of hate is related to 'hate speech'. We started by considering the collocates of "hate" in the journalistic corpus, since the nature of the corpus made it possible to assume that the term would be mostly used in relation to hate speech matters. The most significant and recurrent grammatical pattern associated with "hate" (noun) is its function as a modifier of other nouns, producing expressions such as "hate crime", "hate message" and so on (Table 1), all of them accounting for hate speech-related phenomena.

Table 1 'hate' (noun) as modifier (journalistic corpus)

Grammatical pattern	**Top-ten collocates with frequency and association score**
modified by 'hate' (189/65.6)	crime (94/13.2), incident (16/11.1), message (9/10.2), preacher (7/10.1), law (12/9.71), figure (5/9.53), group (8/9.11), content (3/8.43), literature (2/8.4), panel (2/8.36)

Searches of the pattern in the *enTenTen20* confirmed its association with hate speech. Table 2 features the first 20 collocates, with the frequencies in brackets. The ten most frequent ones are underlined, and the top five are in bold. Collocates are divided by themes, defined after close reading of the frequency lists.[2]

Table 2 'hate' (noun) as modifier (noun) (*enTenTen20*)

Topic	**Collocates (frequency)**
Actions	**hate crime (67459) hate campaign (2693)** hate incident (1340) hate mongering (723) hate violence (626)
Actors	**hate group (14228)** hate monger (1322) hate figure (779) hate preacher (707)
Objects	hate literature (756) hate list (752) hate symbol (503)
Written / spoken	**hate speech (39422) hate mail (7546)** hate propaganda (1260) hate message (1236)
Legal	hate bill (1651) hate law (1351)
Location	hate site (1321)

The pattern identifies actions performing or promoting hate speech (e.g., 'hate crime', 'hate campaign'), actors of hate speech, with a predominance of groups over individuals ('hate group') and written or spoken manifestations (for the purpose of classification, references to 'hate speech' are included in this theme). Although the theme 'legal' only includes two collocates ('hate bill', 'hate law'), both feature among the top-ten collocates, emphasising the legal nature of the concept.

All the expressions metonymically refer to 'hate speech': 'hate', as the defining or most salient characteristic of hate speech, stands for the phenomenon, allowing for the construction of a category of 'hate speech related entities' (DEFINING PROPERTY FOR

[2] The collocate "hate relationship" is not included for not being relevant to the topic; concordance lines showed that it refers to the expression "love hate relationship".

CATEGORY[3] metonymy, Radden and Kövecses 1999). Besides the referential function, the metonymy also bears a negative evaluation of any actor, action or product identified. It is open to question whether 'hate' denotes the emotion of hatred or only some degree of dislike by referring to the extreme, which would involve another metonymy on its own (UPPER END OF SCALE FOR WHOLE SCALE, Radden and Kövecses 1999). What is clear is that the lay understanding of hate speech is indissociable from the negative evaluation attributed to hatred.

The recurrence of the hate metonymy poses the question whether the expressions are discursively used to convey negative evaluations or feelings, or whether they serve referential purposes, without involving extra pragmatic effects. To examine the evaluative functions of the metonymy, we considered the first 200 concordances for the grammatical pattern of nouns modified by "hate" retrieved from *enTenTen20*. Concordances were coded following Martin and White's (2005) evaluation framework, considering only the most basic attitude types. Table 3 summarises the results and the examples reflect blogs, news and discussion genres. The coding "non-applicable" identifies uses that do not refer to hate speech, such as allusions to "love/hate" relationships or the explanation of the card game in the example.

Table 3 Evaluation associated with "hate" + noun collocates

Evaluation	**Total**		**Examples**
Non-applicable	23		• *However this does not mean that the non basic* ***hate cards*** *are a good tool to stop 4 and 5 color decks.*
Appreciation	30	Identification of something as hate speech related	• *We consider this attack not only as an obvious* ***hate incident****, but also as an act of intimidation of activists ...* • *These* [nazis and the KKK] *are* ***hate organizations****, whereas mine is for Christian parents, who are doing the best that they can for their children.*
		evaluation of the hate speech related event or matter	• *"The worst* ***hate speech*** [-Appreciation] *I've heard recently is Richard Di Natale* [-Judgement] *... He's incited violence ...*
Judgement	37	positive (4) of social actors countering hate speech	• *... Tyler Oakley who made a viral video combating* ***hate speech*** *...*
		negative (33) of social actors promoting hate speech	• *he too was charged with inciting violence and* ***hate speech*** • *You, in your extremely simplistic view of the world ... seem to be a* ***hate group*** *dressed in sheep's clothing.*
Affect	3	evaluations of insecurity experienced by the victims	• *Everyday I have a* ***hate crimes*** [sic] *committed on myself.*

[3] Small capitals are used throughout this paper to indicate conceptual status in opposition to linguistic expressions.

Non-evaluative	107		• *We as a community must continue to find ways to end* ***hate crimes*** *and increase funding for programs that work with LGBTQ youth.* • *Of course I don't believe that evolution is the reason behind all* ***hate crimes*** *and racism ...*

Since ‘hate’ is an emotion, Affect (expression of feelings) would be expected. However, expressions of Affect are anecdotal; only three occurrences connoting insecurity of the victims were identified. The main evaluations are Appreciation and Judgement (assessment of things or performances and behaviours), with a predominance of the latter. The negative Appreciations explicitly identify phenomena as hate speech-related or assess hate speech events. In the first case, the evaluation occurs in identifying a particular event with the negative-value laden metonymy –in the example, “attack” is defined as “hate incident” and particular organisations as “hate groups”. In the second case, evaluations are triggered by the qualifiers used to appraise the phenomenon described (“the worst”). Most Judgements are evaluations of social sanction of the actors of hate speech behaviour, except in those cases where the individuals are praised for engaging in countering actions. Negative and positive judgements are based on the negative value associated to the phenomenon, either being countered, in which case individuals are evaluated positively, or perpetrated, triggering negative assessments of the actors.

However, most of the uses of the hate metonymy do not express any evaluation in discourse, beyond the negative valence associated with the concept itself. In those cases, the metonymy is used referentially only, identifying the phenomena as hate speech-related, but without evoking any evaluation of the perpetrators or the actions –in the examples, “hate crimes” is used for denotation purposes only.

The hate metonymy reveals the emotion of hate as central for the conceptualisation of hate speech; it infuses the concept with a negative valence which can be discursively used to evaluate performances, behaviours and social actors related to the phenomenon. The conceptual reliance on ‘hate’ satisfies a primary cognitive function, i.e., identifying hate speech-related phenomena, without necessarily connoting hatred or expressing any evaluation beyond the negative semantics of the concept.

4.2 Extreme negative evaluation as core meaning: promotion of hate speech and countering measures

While metonymic references to hate speech are not necessarily evaluative in discourse, the construal of behaviours promoting, and countering hate speech shows that the concept shares with hate an extreme negative evaluation. Promotion and countering actions were examined studying the collocates of “hate”, “hate speech” and “hatred” in the journalistic corpus with the Word Sketch tool, focusing on those grammatical patterns that identify such behaviours –notably, verbs with the nodes as object to identify actions, and *of* prepositional phrases in pre-modifying position to identify nominalised actions and related entities (Tables 4-5).

Table 4 hate speech promotion and countering actions (journalistic corpus)

Grammatical pattern	Collocates
*Verbs with ‘**hate’** as object* (23/7.99)	spread (5/11), organise (2/10.6), provoke (1/10.2), promote (3/10.1), preach (1/9.95), incite (3/9.88), feel (1/9.83), counter (1/9.48), fight (1/9.38), contain (1/9.38)
*Verbs with **'hate speech'** as object* (334/33)	spread (19/10.5), constitute (13/10.1), combat (13/10), tackle (11/9.74), remove (11/9.59), use (13/9.56), propagate (8/9.42), direct (8/9.36), define (8/9.35), address (8/9.35)
*Verbs with ‘**hatred’** as object* (77/36.8)	incite (22/12,2), promote (10/11,2), normalise (4/10,6), base (4/10), direct (3/9,91), express (3/9,67), cause (3/9,63), spread (3/9,5), cover (2/9,28), stop (2/9,09)

Table 5 hate speech promotion and countering related nominals (journalistic corpus)

Grammatical pattern	Collocates
*Nouns in “of” prepositional phrases that pre-modify ‘**hate’*** (23/7.99)	tsunami (2/11.4), act (3/11.2), ideologue (1/10.4), instigation (1/10.4), teacher (1/10.4), lava (1/10.4), root (1/10.3), extent (1/10.2), element (1/10.2)
*Nouns in “of” prepositional phrases that pre-modify **'hate speech'*** (170/15)	spread (14/11.2), definition (7/10.2), use (7/10.1), instance (6/10.2), guilty (4/9.53), issue (4/9.48), form (4/9.33), producer (3/9.14), perpetrator (3/9.14), prevalence (3/9.12)
*Nouns in “of” prepositional phrases that pre-modify ‘**hatred’*** (23/11)	incitement (6/12,6), advocacy (3/11,9), climate (2/11), stir (1/10,4), emotion (1/10,4), other (1/10,4), lava (1/10,4), politics (1/10,2), element (1/10,2

Although the frequencies of collocates in Tables 4 and 5 are very low due to the small size of the journalistic corpus, they provide some insights into the perception of hate speech reflected in the news. “Hate” and “hate speech” are reported as something expanding across the social sphere (“spread”, “propagate”), and hate is associated with natural forces, as revealed by the collocates for the *of* prepositional phrase in post-modifying position (“tsunami”, “lava”), emphasising the propagation of hate speech and dreadful outcomes that may result from it. Countering actions are portrayed as a battle (“fight”, “combat”), as a competitive sports opposition (“tackle”), or as the containment of hate. The majority of the collocates included in Table 4 emphasise the intensity of hate, hate speech propagation, and countering actions, stressing the negative evaluation associated with the concept.

Collocates of “hate speech” were further examined in the *enTenTen20* for the same grammatical patterns (Table 6), along with the lemmas of natural forces revealed in the journalistic corpus. Sections 4.2.1 and 4.2.2 consider hate speech perpetration and countering actions, respectively.

Table 6 hate speech promotion and countering related collocates (*enTenTen20*)

Grammatical pattern	Top-ten collocates with frequency and association score
Verbs with 'hate speech' as subject (4484/11.2)	incite (28/5,88), spew (13/4,22), victimize (7/3,8), target (118/3,18), fuel (29/2,49), harm (9/2,28), offend (7/2,13), circulate (8/1,32), thrive (9/1,02), spread (23/0,91)

Verbs with 'hate speech' as object (13249/33.2)	spew (122/6,29), criminalize (86/5,84), counter (313/5,47), combat (432/5,09), spout (35/4,83), curb (140/4,81), condone (40/4,66), outlaw (52/4,61), criminalise (12/4,48), propagate (67/4,48)
Nouns in "of" prepositional phrases that pre-modify 'hate speech' (5374/13.5)	purveyor (55/6,21), criminalization (19/5,27), incident (95/4,87), censorship (16/4,83), normalization (15/4,74), prohibition (40/4,66), harm (13/4,6), perpetrator (35/4,58), dissemination (59/4,51), accuse (106/4,5)

4.2.1 Perpetrating hate speech

Hate speech promotion is portrayed as a widespread advancing phenomenon ("spread", "thrive"), and harmful for the targets ("harm", "offend") (Table 6), in line with the pilot results. Verbs describing hate speech behaviour often connote negative judgements of the actors ("spew", "harm", "spout"), stressing the negative evaluation associated with the phenomenon (Table 6). Examining concordance lines made it possible to check the valence of the collocates and their use in relation to hate speech. For example, collocates functioning as direct objects with "fuel" show that the verb emphasises the increase and intensity of the entity referred to, regardless of its valence (e.g., "speculation", "growth", "fire", "demand", "passion"). However, in concordances of "fuel" with "hate speech" as subject, the verb is exclusively associated with negative value laden objects (e.g., "violence", "unrest", "extremism", "atrocities"). Concordances of "hate speech" as subject with "spread", a verb which does not connote a negative evaluation on its own,[4] reflect the intensity and strong negative evaluation associated with the phenomenon:

(1) Hate speech is **spreading easily and very quickly** through phones and social media […] (civicus.org)
(2) "Hate speech is **spreading like wildfire** in social media. We must extinguish it," the Portuguese diplomat said. (news.un.org)
(3) Hate speech is **spreading virally** anyway, with deadly consequences. (theverge.com)
(4) has backed "new forms of self-policing by social media platforms" and action by volunteer groups to fight hate speech **spreading at "lightning speed"** through digital media. (business-standard.com)

These examples show that "spread" acquires a strong negative prosody in portraying the propagation of hate speech, emphasising its rapidity and scale. The rapid speed and extent can be represented literally (example 1), or via conventional metaphorical similes (example 2) and metaphors (examples 3-4). Similes and metaphors evoke natural forces (HATE SPEECH IS A WILDFIRE, example 2, and HATE SPEECH SPREAD IS LIGHTNING example 4), or identify hate speech with a virus (example 3). Similes can establish

[4] A Word Sketch of the common collocates of "spread" (verb) reveals both negative and positive prosodies. The verb collocates with negative value laden nouns ("rumour", "virus", "disease", "lies"), positive ones ("awareness", "love", "joy", "wealth"), and non-evaluative ones ("word", "wings", "legs", "message", "news"). This contrasts with the negative prosody found in other verbs such as "spew", whose direct objects exclusively show a negative valence (e.g., the top five collocates comprise "venom", "lava", "hate", "vitriol", "bile").

metaphorical mappings that are further elaborated with metaphors (example 2, underlined). Metaphorical representations function as evaluative resources, emphasising the negative evaluation associated with hate speech and allowing for stronger appraisals than those evoked by literal qualifications;[5] –compare “spreading easily and very quickly” (example 1) with the hyperbolic metaphorical simile “spreading like wildfire” (example 2). While the reference to lightning in example (4) does not carry a negative evaluation on its own, the negative value is evoked by the ‘fight’ metaphor (underlined), construing hate speech as an enemy advancing at high speed.

These natural forces metaphors resonate with the references to “tsunami” and “lava” in the journalistic corpus, which were also identified among the collocates of “hatred” in the *enTenTen20*, albeit showing in relatively low numbers.[6] References to ‘lava’ and ‘tsunami’, as those in examples (5-6), explicitly identify the natural disasters with the emotion of hate (‘of hatred’ functions as qualifier):

> (5) […] perverse media whose tongues and lips are **like the active volcanoes in Hawaii and Guatemala, spewing out their hot lava of hatred** and lies. (aboverubies.org)
> (6) Unless this is checked, and soon, Pakistan itself may not survive the **tsunami of hatred** that **grips** much of the country. (icit-digital.org)

Natural forces (usually fire-related) and virus-related metaphors stress the rapidity and uncontrollability of hate speech expansion and the devastating consequences for the victims and the social community. Uses of extended metaphors, i.e., metaphors whose linguistic expression extends over two or more clauses (examples 2, 5) (Crisp 2005, 116), reinforce the negative appraisals, occasionally offering elaborate hyperbolic portrayals (example 5). Identifying hate speech with a creature (example 6, “grips”) also depicts the phenomenon as being out of human control. These observations cohere with metaphor research, which has identified natural forces metaphors as recurrent portrayals of intense emotions, emphasising their violent character and uncontrollability (Deignan 1995, 153; Kövecses 2004, 71), and as recurrent representations of crises (Silaški and Đurović 2011). Natural forces metaphors can conceal perpetrators and victims, avoiding any active blaming of the social actors that originated the crises (Silaški and Đurović 2011, 231). In hate speech representations we observe concealment of perpetrators (example 6), but also overt identifications and negative evaluations (example 5).

4.2.2. Fighting hate

Countering actions are consistent with the portrayals of hate speech behaviour. The journalistic corpus featured representations of measures in war and competing sport terms (Table 4), with “combat”, “fight” and “tackle” listed among the verbs collocating with ‘hate’ and ‘hate speech’ as objects. Representations of countering actions as a combat are also recurrent in the *enTenTen20*, with “combat” featuring among the top-ten collocates for those verbs with ‘hate speech’ as object (Table 6). The lemmas “fight”, “combat”, “tackle”, and other fight-related terms such as “battle” and “struggle”, postmodified by the prepositional phrase ‘*against* hate speech’, were searched in the *enTenTen20* to determine whether the WAR metaphors are used beyond

[5] Example 1 provides a literal description of the spread, but attributing ‘hate speech’ the quality to spread on its own is a personification.

[6] The collocates “tsunami of hatred” gave 19 hits, and “lava of hatred” 6 hits.

the journalistic register. Searches also considered the war-related terms as verbs with “hate speech” as object.[7] WAR metaphors were identified in all cases (examples 4 and 7):

> (7) The **battle against hate speech** is a universal and international one, and while geoblocking helps to **protect** people in certain countries from accessing and seeing hateful content, it does not **combat** the issue as a whole. (media-diversity.org)

WAR metaphors or, more generally, FIGHT metaphors, are ubiquitous (Lakoff and Johnson 2003/1980, 4-5) and populate a myriad of discourses such as politics (Musolff 2016, 15-22), health (Semino 2008, 164-166), environment (Atanasova and Koteyko 2015) or business (Charteris-Black 2004, 142-146). Identifying them in descriptions of hate speech measures is thus not surprising. The extended metaphor in example (7) illustrates the pervasiveness and relevance of the WAR metaphor in construing the phenomenon of hate speech, with enemies (perpetrators), victims (targets) and a battle to deploy (regulations). WAR metaphors evoke strong negative evaluations of the perpetrators and their behaviours, and offer a clear plan of action, licensing the enforcement of strong policies, legislations and punishments: as a social enemy, hate speech requires the deployment of strong measures; the governments and legal systems are identified as the defenders of the social community, and the perpetrators as the enemies of social order.

5. Offence and hate speech

Hate and the negative emotions that arise from it are inseparable from the lay understanding of hate speech. These negative emotions are in great measure derived from the impact that hate speech has upon the victims, as illustrated in identifying the phenomenon with natural forces or a social enemy. Characterising hate speech in terms of being hateful and hurting the targets places the issue of distinguishing between offensive and hate speech on the table. Examples (a-b) in section 1 are hateful and offensive, but they make no reference to the social groups of the victims, which following the law is an essential characteristic of hate speech.

Haugh and Sinkeviciute (2019, 198) identify “offence” as a transgression, which involves the transgressive act and the target’s perception of the moral violation involved in causing offence, manifested as hurt feelings. Being offensive is also associated with insulting the target (Haugh and Sinkeviciute 2019, 200). Culpeper and Haugh (2021) characterise offensive speech as insulting, abusive and hurtful, judged as morally wrong and directed towards individuals or groups, mainly characterised in terms of religion and ethnicity. These definitions reflect the difficulty in distinguishing between offensive and hate speech; “abusive” and “insulting” figure in the UK Public Order Act 1986 regulating hate speech, together with religion and ethnicity as targeted attributes.

[7] The grammatical pattern of the prepositional phrase “*against* hate speech” identified 42 hits of “fight” (noun), 21 of “fight” (verb), 23 of “battle” (noun) and 7 of “struggle” (noun). Besides “combat”, the grammatical pattern of verbs with “hate speech” as object also identified 127 hits for “fight”, and 184 hits for “tackle”, portraying countering actions as a competitive sport.

The journalistic corpus and the *enTenTen20* also reflect the association of hate speech with offence. The expressions “offensive content” and “offensive material” featured among the first 100 word-cluster keywords of the journalistic corpus, and “offence” was retrieved as the first collocate for the pattern of nouns modified by “hatred” (frequency: 5, logDice 11.3). Table 6 features “offend” among the top-ten collocates for the grammatical pattern of verbs with “hate speech” as subject in the *enTenTen20*. In this section we argue that, while being associated with offence, hate speech is a narrower concept, which is reflected on the legal dimension of the latter. We continue considering the semantics of the legal descriptors “abusive”, “insulting” and “threatening” to explore whether it is possible to propose further characterisations of hate speech that will help distinguish the two phenomena.

5.1 Hate speech as regulatable speech

The legal nature of hate speech entails that, in contrast to the ‘moral transgression’ identified with offensive speech (Haugh and Sinkeviciute 2019), hate speech also involves a legal transgression. The analysis reflects the legal nature of the phenomenon in both corpora. Table 6 shows that measures against hate speech tend to involve legal procedures (“outlaw”, “criminalise”, “prohibition”). Examining the top-ten collocates for the grammatical pattern of nouns post-modified by the prepositional phrase “*against* hate speech”, which exclusively identifies countering actions, confirms the legal nature of the phenomenon (Table 7).

Table 7 Hate speech countering measures related collocates (*enTenTen20*)

Grammatical pattern	**Top-ten collocates with frequency and association score**
noun post-modified by the prepositional phrase “against 'hate speech'” (800/2.01)	guideline (7/8,05), policy (71/7,72), law (165/7,49), rule (43/7,42), legislate (8/7,27), legislation (11/7,1), Action (7/5,59), prohibition (24/5,54), stance (13/5,49), measure (14/4,52)

Associating hate speech measures with legal sanctions responds to the lay construal of hate speech as a social enemy (section 4.2.2). These observations resonate with Millar’s (2019, 150) identification of the perlocutionary effect of posing a “threat to social peace” as a defining characteristic of hate speech. Importantly, this “threat” does not only involve the harm inflicted on the victims, but also the capability to encourage other individuals to adopt the discriminatory behaviour (Assimakopoulos 2020, 187). The similarities between offensive speech and the legal descriptions of hate speech revealed in the next section provide further support for understanding the perlocutionary effect of incitement as a central characteristic of hate speech from a legal perspective.

5.2 The legal characterisation of hate speech

The UK Public Order Act 1986 characterises hate speech as “threatening”, “abusive” and “insulting”. These descriptors are also used in legal-related texts beyond the UK Public Order. Facebook Community Standards include as ‘hate speech’ contents with “threat of harm”, “offensive and […] insulting labels”, which have “the intent to insult” other users.[8] Abusive behaviour and threats are also mentioned in Twitter's policy on hate speech.[9] Sections 5.2.1 and 5.2.2 examine whether these expressions are exclusively used in legal registers and explore their most salient meanings.

5.2.1 Uses of legal-related terminology beyond legal registers

In legal constructions, it is common for the three descriptors ("threatening", "abusive", "insulting") to appear together with the "and", "or" conjunctions. The top-ten collocates of "threatening" for the grammatical pattern "*threatening* and/or…" feature hate speech-related terms, including "abusive" and "insulting" (Table 10, section 5.2.2). By considering the first 200 concordances for the patterns "threatening and/or abusive" and "threatening and/or insulting" (Tables 8-9), we examined how the expressions relate to hate speech and whether the contexts of use expand beyond the legal register. Concordances were coded as "yes" if the context allowed us to infer a reference to hate speech as defined in the Public Order Act or could potentially entail it, and with "no" otherwise. For the context of use, concordances were tagged as "legal" when they were part of a legal-related text (e.g., social media terms and conditions), as "offence", in those cases where the concordance does not belong to a legal text but describes some offence (e.g., news articles), and as "no" otherwise. The online sources were checked to help coding.

Table 8 Correspondences between references to 'hate speech' and context of use for the grammatical pattern "threatening and/or insulting" (*enTenTen20*)

Hate speech related	**Context of use**		**Examples**
Yes (total: 154)	Legal	79	*You must therefore refrain from making any comments that are defamatory, libellous, **insulting, threatening,** discriminatory, obscene or racist.*
	Offence	47	*A 16-year-old boy was charged under Section 5 of the Public Order Act, for allegedly using **threatening or insulting** words or behaviour.*
	Not legal	28	*I dealt with numerous **insulting and threatening** anonymous letters in my daily correspondence.*
No (total: 46)	Legal	1	*Aggression: Frowning, snarling, baring teeth, staring, with redden face, rigid body, clenched fists, and large, **threatening and insulting** gestures, you display unexpectedly sudden movements* […]
	Offence	3	*On each occasion, the attackers proved to be highly aggressive, **insulting, threatening**, pointing their guns at our staff members and shooting in the air.*
	Not legal	42	*Verbal abuse (yelling, **insulting and threatening**) also often happens in teen relationships.*

[8] 'Hate speech' policy of Facebook Community Standards. Available at: https://transparency.fb.com/en-gb/policies/community-standards/hate-speech/ (accessed 30 May 2022)

[9] 'Hateful conduct policy' of Twitter. Available at: https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy (accessed 30 May 2022)

Table 9 Correspondences between references to ‘hate speech’ and context of use for the grammatical pattern “threatening and/or abusive” (*enTenTen20*)

Hate speech related	**Context of use**		**Examples**
Yes (total: 184)	Legal	144	*Disseminating any unlawful, harassing, libellous,* ***abusive, threatening,*** *harmful, vulgar, obscene, or otherwise objectionable material*
	Offence	30	*In January 2010, Five Luton men were convicted of using* ***threatening, abusive*** *or insulting words or behaviour likely to cause harassment, alarm or distress for abusive chants while*
	Not legal	10	*However, among the numerous* ***threatening and abusive*** *private e-mails Harawira received on the matter* [...] *was one from an individual known*
No (total: 16)	Offence	11	*by his estranged wife of stalking by the sending 760 text messages over a 15-day period. The messages are* ***abusive, threatening****, and in many cases, vulgar.*
	Not legal	5	*She moves into the frame and we watch her* […] *as she tries to deal with her* ***abusive, threatening*** *lover*

Although the descriptors are mainly employed to refer to hate speech in legal registers, they are also adopted in accounts about offences, which may not be hate speech-related (Tables 8-9). Overall, when the patterns are not associated with hate speech, they are used in non-legal contexts. Comparing Tables 8 and 9 suggests that the pattern “threatening and/or insulting” is more commonly adopted to describe potentially hate speech-related language and actions in non-legal contexts, whereas legal registers prefer “threatening and/or abusive”.

Importantly, uses tagged as ‘hate speech-related’ may refer to instances of hate speech, albeit not necessarily exclusively, as observed both in legal (example 8) and non-legal (9) registers:

(8) Rudeness, profanity, **threatening, insulting** posts, personal attacks, defamatory or inflammatory posts will not be tolerated and are a breach of […]

(9) […] after she jumped in front of the car while wearing a white clown suit. She was given a caution for use of **threatening or abusive** words or behaviour or disorderly behaviour likely to cause harassment, alarm or distress.

In example (8), the regulations prohibit hate speech, but we can expect that they also cover posts which threaten or insult an individual, regardless of their adherence to the official hate speech definition. In example (9), it is not clear whether the offender was performing hate speech or simply agitating the public order. These coding difficulties show that, without further specifications, “threatening”, “abusive” and “insulting” allow for rather encompassing descriptions, which may not involve hate speech. While this broad scope may suit social media and institutional regulations, lay uses of legal formulae, such as example (9), raise concerns about the understanding of the descriptors. The fact that these attributes reverberate with official wordings may influence their use and interpretation, potentially leading to misjudgements of the participants involved.

5.2.2 Semantic mapping of "threatening", "abusive" and "insulting"

Examining the most salient meanings of the official descriptors shows how they contribute to the understanding of hate speech. We focused on the collocates for the grammatical patterns: (a) "… and/or …", to identify related characteristics (Table 10), (b) nouns modified by the descriptors, to identify entities considered "threatening", "abusive" or "insulting" (Table 11), and (c) the descriptors complemented by *to* prepositional phrase, to identify usual targets (Table 12).

(a) Characteristics. Although the descriptors tend to appear together in "and/or" patterns (Table 10, in bold), they are associated with very different characteristics. Except "non-life", collocates with "threatening" are hate speech-related, stressing the aspect of hate and the possibility of being hurtful. "Abusive" shares some collocates with "threatening" (underlined); however, while the collocates also connote the potential of harming the victim, they mainly describe individuals' characters and behaviours ("alcoholic", "violent", controlling"). "Insulting" is associated with verbal and non-verbal behaviours that can imply hate speech ("hurtful", "threatening"), it also applies to impolite behaviours ("rude", "dismissive") and it is explicitly related with "offensive".

Table 10 Collocates of "threatening", "abusive" and "insulting" with the "and/or" conjunctions grammatical pattern (*enTenTen20*)

Grammatical pattern	**Collocates with frequency and collocate score**
threatening and/or… (39129/37,1)	non-life (3051/11,2), **abusive** (3629/10,2), obscene (853/8,74), defamatory (524/8,3), intimidating (503/8,14), hateful (449/7,73), libelous (275/7,66), unlawful (413/7,37), **insulting** (274/7,23), harmful (675/7,12)
abusive and/or… (62563/26,9)	**threatening** (3427/10,1), **insulting** (1434/9,15), neglectful (1073/9,07), controlling (1210/8,82), alcoholic (1148/8,62), violent (3244/8,39), deceptive (902/8,28), obscene (700/8,07), defamatory (582/7,96), manipulative (630/7,83)
insulting and/or … (19965/26,4)	**abusive** (1434/9,15), derogatory (322/8,46), rude (856/8,38), condescending (259/8,13), disrespectful (298/8,08), degrading (318/7,87), hurtful (193/7,56), offensive (996/7,51), dismissive (150/7,48), **threatening** (262/7,23)

(b) Entities. The three descriptors are associated with very different entities. "Threatening" mainly collocates with bodily behaviours ("posture", "manner") and communication-related objects ("letter", "voicemail") that can perpetrate hate speech. Collocates with "abusive" mainly describe people related with the victim via family relationships, in their majority male ("husband", "father", "boyfriend"). In contrast, "insulting" is mainly associated with communication-related entities ("remark", "nickname", "comment"), in consonance with its association with impolite behaviour.

Table 11 Nouns modified by "threatening", "abusive" and "insulting" (*enTenTen20*)

Grammatical pattern	Collocates with frequency and collocate score
nouns modified by threatening (56437/55,2)	gesture (786/6,26), letter (3725/6), growl (150/5,86), voicemail (94/5,59), posture (284/5,42), manner (1946/5,4), behaviour (10018/5,29), injury (2037/5,24), email (789/5,13), glare (101/5,05)
nouns modified by abusive (148615/64)	husband (5385/8,52), father (5444/7,75), boyfriend (1496/7,6), priest (3373/73), relationship (17598/7,41), behaviour (4482/7,29), childhood (1207/7,01), stepfather (561/6,9), spouse (956/6,88), behavior (7131/6,82)
nouns modified by insulting (26850/35,5)	Turkishness (217/8,03), Islam (1048/7,45), remark (1190/6,92), sanctity (92/6,49), epithet (129/6,33), nickname (168/6,07), insinuation (45/5,56), caricature (66/5,37), mockery (30/4,74), comment (1208/4,69)

(c) Targets. The most salient targets differ radically for the three descriptors. “Threatening” is mainly associated with abstract social entities (“liberty”, “regime”), whereas the top collocates of “abusive” and “insulting” largely refer to human beings. “Abusive” behaviour is directed to individuals related with the perpetrators, usually via family or professional relationships, and women are identified as prototypical targets. “Insulting” behaviours mainly target groups of individuals identified in terms of race and religion (“Muslim”, “black”), or individual’s “dignity”.

Table 12 “threatening”, “abusive” and “insulting” postmodified by *to* prepositional phrase (*enTenTen20*)

Grammatical pattern	Collocates with frequency and collocate score
threatening' + “to” prepositional phrase (593/0,58)	liberty (14/4,49), regime (20/3,72), stability (9/3,26), survival (13/3,25), neighbor (8/2,86), democracy (10/2,6), interest (24/2,27), order (25/2,25), peace (10/2,22), security (18/2,18)
abusive' + “to” prepositional phrase (2493/1,07)	subordinate (15/6,4), girlfriend (26/6,04), wife (159/5,57), spouse (26/5,34), staff (9/4,97), mother (117/4,82), mom (19/4,82), daughter (48/4,56), servant (10/4,2)
insulting' + “to” prepositional phrase (2331/3,09)	intelligence (253/8,24), intellect (15/6,38), Muslims (21/5,49), monarchy (10/5,37), dignity (16/5,03), black (13/4,85), ego (7/4,66), Christian (22/4,25), minority (12/4,25), Muslim (7/4)

Despite being customarily employed together in hate speech definitions, the three descriptors are associated with different phenomena. “Threatening” evaluates entities perceived as a menace, usually applying to bodily and communicative behaviours. However, salient targets of threatening behaviour are social entities (“democracy”, “liberty”, “regime”). “Abusive” is associated with individual-directed actions, often involving being controlling and neglectful, and the targets prototypically have a relationship with the perpetrators. These relationships are frequently familial or romantic, in which case the typical victims are females, and professional related, in which case the victims are typically subordinate staff. “Insulting” is mainly related to verbal behaviour, frequently offensive and diminishing, and assessed as threatening. Prototypical targets are groups of individuals, or individuals identified in terms of some group belonging, typically defined based on race and religion.

In the light of the most salient meanings associated with the legal descriptors, hate speech can be characterised as verbal or non-verbal behaviour directed to an individual or group of individuals, prototypically defined based on race and religion (albeit women also stand as recurrent targets, stressing the need to include misogynistic behaviours within hate speech). Hate speech behaviours may be perceived as involving “hateful” attitudes, typically displaying impoliteness and causing offence, being “disrespectful”, “rude” and “derogatory”, but will also typically entail some threat and harm for the victims, being “hurtful” and “violent”.

6. Concluding remarks

This paper supports those approaches to hate speech that emphasise the centrality of the emotion of hate for the ordinary understanding of the concept, backing linguistic studies that have associated hate speech with hatefulness (Culpeper 2021). Hate speech behaviours normally connote strong negative evaluations, which may be metaphorically intensified. Identifications of hate speech with natural forces and of the countering actions with a war are habitual, stressing the harm inflicted on the victims and the need for strong measures. ‘Hate speech’ and metonymically related expressions can discursively convey negative evaluations of the individuals and their behaviours. Although the discursive use of these terms does not necessarily appear with the expression of emotion, echoing Brown’s (2017a; 2017b, 577) reluctance to identify hate as a necessary condition for hate speech, the negative valence of the concept derives from identifying ‘hate’ as the most salient or prototypical characteristic of ‘hate speech’. These observations support Brown’s (2017b, 574) identification of negative appraisals as central for the ordinary understanding of ‘hate speech’. The metonymy of ‘hate’ is customarily employed referentially to identify hate speech-related phenomena, and its pervasiveness stresses the ordinary association of hate speech with the expression of hatred, converging with those legal descriptions that identify hatred as its defining characteristic.

The ordinary understanding of hate speech is legal-related (as particularly evidenced by the countering measures) and this is essential: a crucial difference between deeming a speech act as offensive or as hate speech is that, while the first one carries a moral evaluation of the speaker, the latter turns the speaker into an offender. Following the legal philosopher Jeremy Waldron, hate speech laws do not protect from offence, but prevent people, especially those individuals from minority groups, from losing their dignity as social members (Waldron, 2012 chapter 5). Thus, hate speech is a much narrower concept than feeling offence. And yet, distinguishing between offensive and hate speech is not always straightforward.

One possible explanation for this difficulty concerns the official description of hate speech. The legal descriptors “threatening”, “abusive” and “insulting” reflect the harm involved in hate speech behaviour, particularly “threatening” and “insulting”, and the denigration of the victims, particularly “abusive” and “insulting”. However, without further clarifications on the target groups these terms also cover impolite speech that, albeit causing offence, would not qualify as hate speech. The difficulties encountered in the analysis presented in section 5.2 to code the concordances with the legal descriptors as ‘hate speech’ or ‘non-hate speech’ related attest to the blurriness between offensive and hate speech, and the use of the descriptors to account for the two phenomena in different registers. Following the Public Order Act, a crucial factor is the intention of the speaker: threatening, abusive or insulting speech is deemed as regulatable when the acts are “intended or likely to stir up […] hatred”. This characterisation entails the existence of a third party (those bystanders whose emotional response is targeted); thus, ‘hate speech’ regulation applies to *public* speech acts.

Following from the legal characterisation of hate speech as ‘*incitement* to hatred’, Assimakopoulos (2020, 187) distinguishes the ‘speaker-intended perlocutionary effects’, i.e., to promote discriminatory behaviours, as the touchstone to distinguish prosecutable hate speech from offence. Identifying hate speech as ‘incitement’ places speaker’s intent at the centre of inquiry. Assimakopoulos (2020, 190) bases intent identification on the Gricean reasoning to recover conversational implicatures; those cases where the Gricean reasoning does not apply would be considered unintentional, hence not accounted as hate speech. However, applying the Gricean reasoning may not always be straightforward. One way of circumventing this difficulty is to identify incitement with hortative or imperative constructions, as in Culpeper et al.’s (2017). Nevertheless, circumscribing incitement to its paradigmatic realizations does not free us from ambiguous cases (e.g., imperatives that do not involve an action causing physical harm) (Culpeper et al., 2017), and risks leaving out those expressions that, while not adopting the paradigmatic forms, contextually function as hate speech instigators.

Computer mediated communication (CMC) such as examples (a) and (b) (section 1) is frequently public or easily accessible, obscuring the distinction between private and public spheres and, by extension, offensive and hate speech. The prolific use of the hate metonymy identified in the analysis, with expressions such as “hate email” or “hate message”, echoes the distortion between the private and the public, and what O’Driscoll (2013, 380) has described as the “public-ization” of private offence: when the private becomes public, “insulting attributes become inadmissible” (O’Driscoll 2013, 379), turning into material for institutional sanction. O’Driscoll (2013) has argued for a redefinition of the private/public distinction which considers the nuances afforded by the new modes of communication. An important implication of CMC is that the actual targets and speakers’ intent may not always be easily identifiable in digital contexts. Hardaker and McGlashan (2016) show how the addressees of hate tweets may not only be the targeted victims, but also the sympathizer community of users that are encouraged to engage in the abusive behaviour.

Driven by practical interests, companies and institutions may opt for all-encompassing definitions such as the examples in section 5.2, deeming ‘public-ized’ offences matters of institutional sanction for the preservation of the harmony and well-being of all workers or users. However, a distinction between offensive and hate speech should nonetheless be provided to account for the legal dimension of the latter. Following the preceding analysis, hate speech is a complex multifaceted phenomenon and, while the expression of hatred constitutes one of its central characteristics, hate alone is not enough to evaluate a certain behaviour as hate speech. The expression of threats and the denigration of the targeted individuals based on a protected characteristic (race, religion and sexual orientation, as specified in the Public Order, but also gender) are essential, together with the potential to cause harm to the victims and the intention of the speaker to stir up hatred.

This paper has aimed to contribute to the study of hate speech and the discussions about the ordinary understanding of the concept. However, this analysis is limited by the nature of the corpora studied; it cannot account for the actual judgements of lay speakers on offensive and hate speech, nor can it determine whether the “public-ization” of private offence has a chilling effect on the community. Considerations such as those demonstrate the need for experimental studies on the public understanding of offensive and hate speech, alongside the kind of corpus analysis undertaken here.

Acknowledgements

I am indebted to Nat Hansen, Emma Borg and Maxime Lepoutre for their support, discussions and feedback, and making this study possible. I am also grateful to the two anonymous reviewers and the editors for their insightful comments.

References

Assimakopoulos, Stavros. 2020. “Incitement to Discriminatory Hatred, Illocution and Perlocution.” *Pragmatics and Society* 11(2): 177-195. doi.org/10.1075/ps.18071.ass

Atanasova, Dimitrinka, and Nelia Koteyko. 2015. “Metaphors in Guardian Online and Mail Online Opinion-page Content on Climate Change: War, Religion, and Politics.” *Environmental Communication. A Journal of Nature and Culture.* 11(4): 452-469. doi.org/10.1080/17524032.2015.1024705

Basile, Valerio, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. “Semeval-2019 Task 5: Multilingual Detection of Hate Speech against Immigrants and Women in Twitter.” In *13th International Workshop on Semantic Evaluation* (pp. 54-63). Association for Computational Linguistics. https://aclanthology.org/S19-2007.pdf

Brown, Alexander. 2015. *Hate Speech Law*. Abingdon: Routledge.

Brown, Alexander. 2017a. “What is Hate Speech? Part 1: The Myth of Hate.” *Law and Philosophy*, 36(4): 419-468. doi.org/10.1007/s10982-017-9297-1

Brown, Alexander. 2017b. “What is Hate Speech? Part 2: Family Resemblances.” *Law and Philosophy* 36:561–613. doi.org/10.1007/s10982-017-9300-x

Butler, Judith. (1997) 2021. *Excitable Speech. A Politics of the Performative.* London/New York: Routledge.

Charteris-Black, Jonathan. 2004. *Corpus Approaches to Critical Metaphor Analysis,* Basingstoke: Palgrave-Macmillan

Crisp, Peter. 2005. Allegory, blending, and possible situations. *Metaphor and Symbol*, 20(2):115-131. DOI: 10.1207/s15327868ms2002_2

Culpeper, Jonathan. 2021. “Impoliteness and Hate Speech: Compare and Contrast.” *Journal of Pragmatics*, 179:4-11. doi.org/10.1016/j.pragma.2021.04.019

Culpeper, Jonathan, and Michael Haugh. 2021. “The Metalinguistics of Offence in (British) English: A Corpus-based Metapragmatic Approach.” *Journal of*

Language Aggression and Conflict, 9(2): 185-214. doi.org/10.1075/jlac.00035.cul

Culpeper, Jonathan, Paul Iganski and Abe Sweiry. 2017. “Linguistic Impoliteness and Religiously Aggravated Hate Crime in England and Wales.” *Journal of Language Aggression and Conflict*, *5*(1), 1-29. doi.org/10.1075/jlac.5.1.01cul

Deignan, Alice. 1995. *English Guides 7: Metaphor. Helping Learners with real English.* London: Harper Collins.

Delgado, Richard. 1982. “Words That Wound: A Tort Action for Racial Insults, Epithets, and Name-Calling.” *Harvard Civil Rights-Civil Liberties Law Review* 17(1): 133-182 https://heinonline.org/HOL/P?h=hein.journals/hcrcl17&i=141

Fairclough, Norman. 1989. *Language and Power*. London: Longman.

Fowler, Roger. 1991. *Language in the News: Discourse and Ideology in the Press.* London: Routledge.

Hardaker, Claire, and McGlashan, Mark. 2016. “‘Real Men don’t Hate Women’: Twitter Rape Threats and Group Identity.” *Journal of Pragmatics*, 91:80-93. doi.org/10.1016/j.pragma.2015.11.005

Haugh, Michael. 2016. “The Role of English as a Scientific Metalanguage for Research in Pragmatics: Reflections on the Metapragmatics of ‘Politeness’ in Japanese.” *East Asian Pragmatics* 1(1): 39e71.

Haugh, Michael, and Valeria Sinkeviciute. 2019. “Offence and conflict talk”. In *The Routledge Handbook of Language in Conflict,* ed. by Matthew Evans, Lesley Jeffries and Jim O’Driscoll, 196-214. Oxon: Routledge.

Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. “The Sketch Engine: Ten Years on.” *Lexicography*, 1(1): 7-36. doi:10.1007/s40607-014-0009-9

Kövecses, Zoltán. 2004. *Metaphor and Emotion. Language, Culture and Body in Human Feeling.* Cambridge: Cambridge University Press.

Lee, Thomas R. and Stephen C. Mouritsen. 2018. “Judging Ordinary Meaning.” *YALE Law Journal.* 127(4): 788-829. https://www.jstor.org/stable/45097958

Lee, Thomas R. and Stephen C. Mouritsen. 2020. “The Corpus and the Critics.” *The University of Chicago Law Review*, 88(2): 275-366. https://www.jstor.org/stable/26986409

Manning, John F. 2002. “The Absurdity Doctrine.” *Harvard Law Review*, 116(8): 2387-2486. https://doi.org/10.2307/1342768

Martin, James R. and Peter R.R. White. 2005. *The Language of Evaluation. Appraisal in English.* New York: Palgrave.

Matsuda, Mari J. 1989. “Public Response to Racist Speech: Considering the Victim's Story.” *Michigan Law Review*, *87*(8): 2320-2381. https://repository.law.umich.edu/mlr/vol87/iss8/8

Mouritsen, Stephen C. 2011. “Hard Cases and Hard Data: Assessing Corpus Linguistics as an Empirical Path to Plain Meaning.” *Columbia Science and Technology Law Review.* 13(1): 156-202. https://heinonline.org/HOL/P?h=hein.journals/cstlr13&i=156

Musolff, Andreas. 2016. *Political Metaphor Scenarios. Discourse and Scenarios*. London/NY: Bloomsbury.

O'Driscoll, Jim. 2013. Situational transformations: The offensive-izing of an email message and the public-ization of offensiveness. *Pragmatics and Society, 4*(3): 369-387. doi.org/10.1075/ps.4.3.05odr

Radden, Günter, and Zoltán Kövecses, Z. 1999. “Towards a Theory of Metonymy.” In *Metonymy in Language and Thought*, ed. by Klaus-Uwe Panther and Günter Radden, 17-60. Amsterdam/Philadelphia: John Benjamins.

Rikhof, Joseph. 2005. “Hate Speech and International Criminal Law: The Mugesera Decision by the Supreme Court of Canada.” *Journal of International Criminal Justice*, 3(5): 1121-1133. doi.org/10.1093/jicj/mqi082

Rychlý, Pavel. 2008. “A Lexicographer-Friendly Association Score.” *Proceedings of Recent Advances in Slavonic Natural Language Processing* (pp. 6-9) RASLAN. https://nlp.fi.muni.cz/raslan/2008/raslan08.pdf#page=14

Sanguinetti, Manuela, Fabio Poletto, Cristina Bosco, Viviana Patti and Marco Stranisci. 2018. “An Italian Twitter Corpus of Hate Speech against Immigrants.” In *Proceedings of the eleventh International Conference on Language Resources and Evaluation* (LREC 2018) (2798- 2805). https://aclanthology.org/L18-1443.pdf

Schmidt, Anna and Michael Wiegand. 2017. “A Survey on Hate Speech Detection Using Natural Language Processing.” In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, April 3-7 2017, Valencia, Spain* (pp. 1-10). Association for Computational Linguistics. https://aclanthology.org/W17-1101.pdf

Semino, Elena. 2008. *Metaphor in Discourse*. Cambridge: Cambridge University Press.

Silaški, Nadežda and Tatjana Đurović. 2011. The NATURAL FORCE metaphor in the conceptualisation of the global financial crisis in English and Serbian. *Zbornik Matice srpske za filologiju i lingvistiku*, 54(1): 227-245.

Solan, Lawrence M., and Tammy Gales. 2016. “Finding Ordinary Meaning in Law: The Judge, the Dictionary or the Corpus?” *International Journal of Legal Discourse*, 1(2): 253-276. doi.org/10.1515/ijld-2016-0016

Taylor, Charlotte. 2015. “Beyond sarcasm: The Metalanguage and Structures of Mock Politeness.” *Journal of Pragmatics*, *87*: 127-141. doi.org/10.1016/j.pragma.2015.08.005

Taylor, Charlotte. 2017. “The Relationship between Irony and Sarcasm: Insights from a First-order Metalanguage Investigation.” *Journal of Politeness Research*, 13(2): 209-241. doi.org/10.1515/pr-2015-0037

Tobia, Kevin, Brian Slocum, and Victoria Nourse. 2023. “Ordinary Meaning and Ordinary People.” *University of Pennsylvania Law Review* 171: 1-14.

Waldron, Jeremy. 2012. *The harm in hate speech*. Cambridge/London: Harvard University Press.

Warner, William and Julia Hirschberg. 2012. “Detecting Hate Speech on the World Wide Web.” In *Proceedings of the Second Workshop on Language in Social Media, June 7 2012 Montréal, Canada* (pp. 19-26). Association for Computational Linguistics. https://aclanthology.org/W12-2103.pdf

University Park, Nottingham, NG7 2RD, United Kingdom.
sara.vilar-lluch@nottingham.ac.uk | s.vilar.lluch@gmail.com
ORCID: 0000-0002-5495-9386