Studii de lingvistică 14, nr.2, 2024, 39 - 62

Annotating linguistic features of extreme narratives: a literature review and a proposal

Fabienne Baider[1]
Alexandros Gregoriou[2]

Abstract: The research questions addressed in this paper are twofold: 1. How can we define an extremist narrative? 2. Can a narrative's toxicity be determined according to certain linguistic features? To this end, we first define what is currently considered an extreme narrative, with a focus on identifying the differences between hate speech and extreme speech. We next review annotations of hate speech and extreme speech, with a specific emphasis on linguistic features; then, drawing on our results of that review we propose an annotation schema to label extreme speech. Finally, we share the results of a small sample we annotated and reveal the most frequently and most consistently annotated features that could be representative markers of extreme speech.

Keywords: extremist narratives, hate speech, annotation schema, toxicity, linguistic features

1. Defining fundamental concepts of the study

1.1. What is data annotation?

Data annotations are integral to both supervised machine learning and deep learning algorithms. They define the features of specific utterances in a way that enables us to build an algorithm that can make accurate predictions. For example, the most frequent defining features of hate speech are: 1. the speaker / writer's *intent to incite the audience to do harm* against a targeted group; 2. advocating *violence or hatred*; 3. *the group is historically disadvantaged* and

[1] University of Cyprus, French and European Studies Department; baider.fabienne@ucy.ac.cy.
[2] University of Cyprus, French and European Studies Department; gregoriou.alexandros@ucy.ac.cy.

vulnerable (the Council of Europe Framework Decision 2008). Hence, we can consider the following expressions – *incite to do harm*, *advocating violence or hatred* and *the target being a specific group* – to represent the parameters of hate speech when annotating utterances in a corpus. These would belong in what is called a 'schema,' i.e., all the parameters considered when working on the data. While the above features define hate speech semantically, other grammatical features might also be recorded as typical elements of hate speech, for example, the use of the pronouns *we* or *them* (Ascone and Longhi 2018; Machado Carneiro *et al.* 2023).

Manual annotation generally precedes automatic annotation. As such, the role of the annotators is the key for the validity of the annotation schema. Typically, annotators follow a multi-stage training program to ensure maximum consistency and coverage. Precise guidelines and multiple examples are used to instruct annotators, and because annotator (dis)agreement is a common challenge, their responses must be closely monitored. Whenever a statistically significant disagreement is recorded, a discussion will follow to resolve the conflict and to examine for any bias that might undermine the homogeneity of the annotation results (Carvalho *et al.* 2022; Yuan and Rizoiu 2025). Indeed, although linguistic features are, in theory, transparent, they can be differently interpreted due to underlying human bias, such as a certain class perspective or 'normative expectations' (Baider 2020; Postmes *et al.* 2000). Maronikolakis *et al.* (2022) noted, for example, that the sociological profile should be diverse since in a 2018 study carried out by Founta *et al.* (2018) 66% of the annotators were male, while in Sap *et al.*'s (2020) research, 82% were white.

Each of the various social media giants, such as Facebook, has their own complex set of rules to annotate hate speech. Examples of the complexity of these rules (Baider 2023) are given below. For instance, Facebook asks their evaluators to identify subtle differences, e.g., *Migrants are so filthy* (non-violating - ignore) vs *All English people are dirty* (violating - delete); or *fucking migrants* (non-violating - ignore) vs *fucking Muslims* (violating - delete). The reasons for these different decisions (ignore or delete) are the following:

- migrants are only a "quasi-protected category" and therefore expressing disgust against them is allowed under certain circumstances. If statements such as "migrants are filthy" are allowed, the statement "migrants are filth" is deleted. The latter refers to a well-established racist metaphor <Migrants are DIRT>;
- *All English people are dirty* is deleted, as condemning people based on their nationality is not allowed. This rule is consistent with the Council of Europe definition specifying nationality

among the criteria that must be fulfilled for a comment to be labeled as hate speech.

Excel spreadsheets can be used for annotating data; below (Figure 1) is an example of the Excel spreadsheet used to tag triggers of hate speech (Baider and Romain 2022):

Triggered by	Trigger of hate speech						Triggered comment N°
	TOPICS		RHETORICAL MODE		THREAT		
	Immigration			display of negative emotions			#9, 15, 38, 49, 51, 57, 71, 75
8	Immigration		history				#10
9						morality (dishonest)	
	Media &Public personae			sarcasm		morality (dishonest)	#12
11	Immigration			use of swear words	Social	morality (dishonest)	#13
12							
	Immigration		history				#16
8	Politics/Ideology					cognitive abilities (reasoning)	
14	Immigration		history	use of swear words	Social		#18
	Immigration		facts				#20
16	Politics/Ideology			personal attacks		cognitive abilities (reasoning)	#21

Figure 1: Example of annotating hate speech

Data must be correctly structured and labeled for the machine learning systems to use it to perform given tasks. Once the data are annotated, the resulting datasets can be the basis for creating and training models for machine learning.

However, we must bear in mind the limitations of any annotation process and automatic detection. The limitations that were identified by Fortuna and Nunes (2018: 25, El Sherief *et al.* (2018) in their respective reviews of automatic hate speech detection are:

- The challenge of addressing the low rate of agreement among individuals classifying hate speech; this suggests that the classification process could be even harder for machines (Fortuna and Nunes 2018);
- The use (by authors of hate speech) of misspellings and abbreviations to avoid classifiers;
- The keywords used in machine learning can be used in benign as well as hateful contexts, or in metalinguistic utterances (El Sherief *et al.* 2018; Baider 2020);
- The task requires expertise in culture and social structure to contextualise any comment, since the interpretation or severity of hate terms can vary based on community tolerance and contextual attributes (El Sherief *et al.* 2018; Fortuna and Nunes 2018; Baider 2020);
- The evolution of social phenomena and the creativity of language make it difficult to track all racial and minority insults (Fortuna and Nunes 2018);
- Despite the offensive nature of hate speech, abusive language

can often escape precise definitions, e.g., the use of sarcasm is common (Baider and Constantinou 2020);
- Hate speech detection is more complex than simple keyword detection (El Shereif *et al.* 2021).

We will discuss these limitations more fully in the next section.

1.2. What is extreme speech?

This article is partially based on our preliminary research work carried out within the ARENAS project[3] focused on extreme narratives. To annotate data, it is necessary to have a very clear definition of the main concepts to be identified in discourse excerpts such as Tweets or Facebook posts (Davidson *et al.* 2017; Sap *et al.* 2020; Wich *et al.* 2020). In fact, while there have been many studies annotating hate speech (Machado Carneiro *et al.* 2023, for a recent review; Fortuna and Nunes 2018), this is not the case for the topic of extreme speech, which has been much less studied. Therefore, we pose the questions: how are the two concepts similar and what is it that differentiates them? To answer these two questions, we use the research of Pohjonen and Udupa (2017) and that of Maronikolakis *et al.* (2022), as well as the categories identified in our previous research (Baider 2020, Baider 2023).

1.2.1. Features common to extreme speech and hate speech

Regarding the definition of extreme speech, as we noted above, hate speech and extreme speech are very closely related. First, both comprise utterances that transgress the accepted norms in a given community at a given time. Second, neither concept has a universally accepted definition (Guillén-Nieto 2023; Baider 2023). Third, in both cases societal, cultural and historical contexts must be considered when deciding if an utterance reflects extreme or hate speech within a specific community (Fortuna and Nunes 2018; Baider 2020). This need for contextualization is especially true for hate speech, as there can be legal consequences if it is proven that an utterance is hate speech (Waldron 2012; Langton 2018; Alkiviadou 2019; Baider 2023).

[3] The ARENAS project (Analysis of and Responses to Extremist Narratives) is co-funded by the European Union's Rights, Equality and Citizenship Programme (2023-2027). The project ARENAS grant no. 101094731 is available at http://arenasproject.eu/. The European Commission support for the production of this publication does not constitute an endorsement of the contents, which reflect the views of the authors alone, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

The plurality of hate speech definitions has been noted in research targeting hate speech from a sociolinguistic perspective (Baider 2020), and this is also true for extreme speech: evaluators of extreme/hate speech data are well aware of the difficulty in deciphering meaning when they are in the position of being an outsider to the situational context (O'Sullivan and Flanagin 2003: 73). These authors also note that "intentional uses of non-normative language have diverse specific interactional goals" (*ibid.*). Moreover, in some communities, the normative language might include or even require using of abusive language and profanity (Baider 2020). In the same way, Guillén-Nieto (2023), in a review of current legal approaches to hate speech, concludes that pragmatic theory, which considers the specific situational context, improves our understanding of the law's wording when assessing utterances for hate speech.

1.2.2. Differences between hate speech and extreme speech

In terms of differences, as early as the 1990s, Walther *et al.* (1994: 477) argued that the "social dynamics of the new media" had to be considered before labeling violent or uncivil interaction as hate speech. In other words, while although an utterance can be extreme, it cannot be considered hate speech unless it fulfils the three essential criteria at all times: 1. the speaker / writer's intent *to incite the audience to do harm* against a targeted group; 2. the speaker / writer's message incites to *violence or hatred*; 3. *the target group is historically disadvantaged and vulnerable*. Extreme speech would not fulfill these three criteria at the same time, as we explain below.

Some authors, including Udupa and Pohjonen (2019) and Udupa *et al.* (2021), have defined extreme speech as speech "that pushes the boundaries of civil language". It follows, therefore, that the social and historical contexts must be understood for their role in molding such speech. However, like for hate speech, the societal, cultural and historical backgrounds will determine what can be considered to be "beyond the boundaries of civil language" (Maronikolakis *et al.* 2022).

Other researchers, e.g. Udupa *et al.* (2021), differentiate between extreme speech and hate speech according to the specific approach used to study these utterances: extreme speech is hate speech studied from an *anthropological perspective*, while hate speech is studied primarily *from a legal or linguistic perspective*. This anthropological perspective highlights the importance of labelling an utterance as 'extreme' by observing the *reactions* of media users to the utterance; it would be these reactions that form the basis for labelling speech as extreme, "vilifying, polarizing, or lethal" (Pohjonen

and Udupa 2017: 1174). The label 'extreme speech' would thus be determined by the reactions/interactions among users; it would not be determined by a preexisting definition or even a preconception of what should be considered as extreme speech. This difference seems to be the perlocutionary dimension asserted by Guillén-Nieto (2023) in her review of hate speech court cases.

In an attempt to define extreme speech, Udupa *et al.* (2021) identified three types of extreme speech, i.e., Derogatory Extreme Speech, Exclusionary Extreme Speech and Dangerous Extreme Speech. Derogatory Extreme Speech insults or belittles its target. The utterances not only cross the boundaries of civility but are also associated with the attitude of contempt and disrespect for their target. However, neither contemptuous nor disrespectful remarks are illegal, even though they are uncivil. They may be considered as "socially unacceptable discourse" (Fišer *et al.* 2017), depending on the cultural and situational contexts. Exclusionary Extreme Speech is based on the phenomenon of polarization and calls for the exclusion of vulnerable groups based on protected attributes (for example, ethnicity, religion and gender). Udupa *et al.* (2021) suggest that this type of speech should require removal just like hate speech. However, even if these exclusionary practices are known to be the first step towards hate speech (Timmerman 2008; Baider 2020), they do not constitute hate speech as such since they do not explicitly call for violence or hatred. Finally, Dangerous Extreme Speech refers clearly to the hate speech definition found in most legal texts; the criteria delineated by Udupa *et al.* (2021) are very similar to the criteria contained in the test put forward by the Rabat Plan of Action (OHCHR 2013). This stratification of extreme speech implies that some types of extreme speech are socially unacceptable, but they are legal since there is no call for violence or hatred. We would agree with only that category to be labelled extreme speech because extreme speech, in our understanding, does not have to target a vulnerable community, which covert and overt hate speech do (Baider 2022; El Shereif *et al.* 2021). The two other types (Exclusionary and Dangerous speech) seem to refer to the existing labels 'covert' and 'overt' hate speech.

In a summary of this overview, we can identify three different criteria to use when annotating extreme speech: presence of extremely negative emotions such as expression of fear or grievances, words leading to polarization and / or exclusionary language. The presence of vulnerable groups in the narrative is debatable as mentioned earlier. A speech can still be extreme, but not address specific communities. In our approach, extreme speech consists, therefore, in utterances considered by the users and in context, as uncivil, but legally acceptable, whilst overt hate speech is confined to violent, uncivil and legally unacceptable speech.

2. How to annotate extremist narratives?

2.1 Annotating the structure of a narrative: a literature review

The ARENAS project's objective was to approach the annotation of extreme speech as a *narrative*. For the concept of narrative, we reviewed structuralist works focused on the grammar of narratives, since they are well adapted to the binary classifications often used in annotation schemes. We considered what Greimas (Greimas and Courtés 1979) and Propp (1928) defined as the minimal units in a narrative structure (Trifonas 2015), and we now summarize what we have chosen to use or adapt for our schema after our review of these studies.

The minimal units making up a narrative are its different actors (actants or agents), items (or objects), and incidents. Propp (1928: 21) identified the minimal unit of narrative analysis to be the *function* in terms of an action. The agents in a narrative are then labelled according to their function in that narrative. Propp identified seven main characters performing these functions (the villain, the donor, the helper, the sought-for person, the dispatcher, the hero, and the false hero), whereas Greimas (1966) identified six functions: subject *vs* object, sender *vs* receiver, and helper *vs* opponent. We chose to use Greimas' schema as they are based on binary oppositions, and we found they could be more easily annotated.

Greimas further posited that narrative sequences always consider a subject and an object at an initial event. This event is followed by a disjunction between the subject and the object that develops first into a *problem*, and then into a (targeted) *final stage*. We can generalize the 'disjunction' sequence or problem as being *the initial event*, whereas the final stage could be labeled as *consequence.* The word *consequence* is important since understanding the causal and consequential relationship between the agents and their actions in extremist narrative is crucial to determining solutions and counter actions to prevent future escalation of violence. Ricoeur (1984) also argued that causality was a basic dimension of the narrative process. Both Greimas and Propp also considered time to be an important dimension, reminiscent of the parameter *setting* as found in Hymes' (1974) SPEAKING model, which encompasses all the spatial and temporal dimensions of a communicative event.

Finally, through a review of the techniques used for manual annotation and analysis of narratives at the level of the macrostructure, we explored the compatibility of Gillam *et al.*'s (2017) Monitoring Indicators of Scholarly Language (MISL) tool for the purposes of annotating extremist narratives. The macrostructure subscale of the tool included some elements resembling the units we had identified

in the research discussed above, such as *character, setting, initiating event* and *consequence*. Despite being originally used in a different line of research (assessing the quality and complexity of children's narratives), the proposed scale seemed appropriate for annotating the units in extremist narratives, and, in fact, it had already been tried with machine learning methods with promising accuracy in automated scoring (Jones *et al.* 2019). Furthermore, it includes additional parameters such as *internal response* and *plan*, which were not noted in other studies we had considered. Based on this short literature review, we proposed the main narrative structure parameters to be annotated by integrating those elements from the MISL which were in line with the functional, causal, and temporal units we identified in the works of Ricoeur, Propp and Greimas with the addition of *helper* and *opponent* based on the work of the latter. We will provide a detailed discussion on the use of the scale in the Annotation Schema section following.

2.2 Annotating the text of extremist narratives

2.2.1. Previous annotations with machine learning

A number of studies in computational sciences reviewed different machine-learning algorithms and techniques for effective detection of hate speech (e.g., Bansal *et al.* 2022; Das *et al.* 2021; Omran *et al.* 2023; Yerden and Turgut 2024). Most of these studies concluded that the classifiers used today are able to distinguish between hate speech and offensive language, as well as between offensive and non-offensive content. However, they mitigate these conclusions by acknowledging the dangers of generalizing classification models to diverse datasets and domains, citing the need to adapt models to the ongoing evolution of online communication platforms (as pointed out earlier by Fortuna and Nunes (2018)).

Adding contextual information is cited as a potential solution to enhance model performance. Omran *et al.* (2023: 9) notes that this information is limited to "user demographics, social network structures, and temporal dynamics". This caveat acknowledges the pervasive fluidity inherent to the task of defining hate speech, but it still does not factor in the linguistic and/or the social context of the utterances (Gagliardone *et al.* 2014; Baider 2020; Sap *et al.* 2020; Dixon *et al.* 2018). Hate speech detections/algorithms may be based on communication models that favor a certain idea/standard of 'decent' communication, which, at the same time, may stigmatize other registers and communities of practice. Indeed, in terms of understanding meaning, the current automatic detection models are unable to perceive a vital element: the illocutionary force, its

perlocutionary dimension, and the context of each utterance, which "must be considered when analyzing its exclusionary force" (Wodak 2015: 207).

Socially unacceptable discourse (SUD) which, we suggest, is often a characteristic of extreme speech, was described by Machado Carneiro *et al.* (2023) in their recent research on SUD as: *abusive, aggressive, hate only identity, hate insult, obscene, offensive, profane, severe, toxic, threat*. They identified only two categories of such SUD that are systematically and consistently annotated: *Abusive speech and Aggressive speech,* both categories fitting the label extreme speech in our approach.

2.2.2. Linguistic features of extreme speech

Various linguists (Musolff 2015; Ascone and Longhi 2018; Baider 2020; Knoblock 2023) have identified the lexical and grammatical means commonly used to express the social processes involved in hate speech. These are based on what van Dijk referred to as "the ideological square" (1998: 33). This square is the basis of the polarization phenomenon described as exclusionary extreme speech by Udupa *et al.* (2021). According to this ideological square, 'our' good properties/actions are emphasized, while 'their' bad properties/ actions are highlighted; this square has also been cited as fundamental to prejudices (Allport 1954). In the paragraph below we emphasize in italics the elements that are most useful for annotating the linguistic features of extreme speech.

As noted by Udupa *et al.* (2021), one primary characteristic of extreme speech involves creating an *in-group/out-group dichotomy,* where differences (real or imagined) are emphasized and similarities are minimized or ignored. This polarity is then constructed in dangerous extremist speech as negative/positive, employing humiliating and contemptuous language, with the use of *insults, slurs, negative stereotypes* and *group defamation*, among other devices (Chakraborti 2015; Brown 2018: 307; Baider 2019).

In parallel, the idea that the out-group poses a *physical, economic,* and /or *social threat* to the in-group is a common feature in discriminatory speech studies (Stephan and Stephan 2000), and this was also identified by Udupa *et al.* (2012) as a characteristic of Dangerous Extreme Speech. *Threat* is a common feature of conspiracy theories, such as the Great Replacement (Baider 2022; Wodak and Richardson 2022), which suggests that white people in Europe will be supplanted by non-white immigrants as part of an orchestrated plan by elites. The threat can be directed at either the target or the speaker (victimization of the speaker). The fear instigated by the supposed threat can lead to calls for annihilation with *imperatives* such as

those used in the manifesto of Brenton Tarrant, the perpetrator of the Christchurch Mosque shootings in New Zealand (*kill, destroy, slaughter, drown the boats*, etc.).

Antagonism and mistrust are encouraged through the use of dehumanizing symbolic language, i.e., the use of metaphors, and especially animal metaphors such as *cockroach* (Timmerman 2008) and *vermin* (Baider and Kopytowska 2018; Musolff 2015) or other non-human metaphors (*Untermensch, dirt, filth, tsunami*). This dehumanization also encourages the use of physical violence.

El Sherief *et al.* (2018) have identified several other linguistic features of hate speech likely to be found in extreme speech:

- the presence of religion (*Jihadis, extermination, Zionazi, Muzzie*), which seems to be a very fertile topic for extreme speech hashtags, also provided useful keywords, including those related to white supremacists (e.g. #*whitepower*) or hashtags targeting specific communities (e.g. #*nomuslimrefugees*);
- the overuse of certain pronouns, namely, *they* and not *we,* and all the 3rd person pronouns and adjectives;
- the dominance of *anger* as the most common emotion found in comments, which indicates the need to identify emotions in our data;
- the *present tense* is more frequent in hate speech than in 'unbiased' Tweets;
- the semantic and lexical field of *death* is more common than in ordinary Tweets, which is what we regard as extreme speech, since it includes irony, white grievance, inferiority language, social stereotypes, threats and misinformation. However, in this research these features apply only to vulnerable communities.

In summary, the linguistic parameters best suited to annotate extreme speech might include:

Use of *insults, slurs, negative stereotypes, defamatory statements,* which leads to polarizing speech
Use of conspiracy theories and other *threat*-provoking arguments or devices
Presence of *negative emotions,* including emoticons and other formal markers (e.g., exclamation marks) and including frustration and emotional grievances
Presence of third person *pronouns*
Dehumanizing devices such as *metaphors, comparisons*
Modality, such as the imperative mood or the obligation mood (*must*, *should*) to incite action and to instill culpability for not acting
Presence of lexical fields pertaining to *religion*

Table 1: Linguistic parameters for extreme speech annotation

These linguistic findings informed the development of our textual annotation schema.

3. Creating our annotation schema

This section describes the small pilot study we needed to carry out before presenting the schema we suggested to the teams of the ARENAS project.

3.1. Main features of extreme speech

Based on both the common parameters identified by other researchers for task 2.2 of the ARENAS project, and the most important concepts we had previously mapped out during our literature review on textual annotation, as seen in Table 2 below, we selected the most likely parameters to include in our schema.

We determined that at least one of the labels – *hostility*, *instillment of fear to out-group, incitement to violence* – had to apply in conjunction with the rest of the categories for a case to be considered a *dangerous* extremist post, but not necessarily extremist speech. At the same time, hostility and expression of extreme frustration (whether through metaphors, slurs, etc.) in itself cannot suffice to characterize content as both dangerous and extremist.

Task 2.2 Definition	**Parameters for the Schema**
clearly distinguish between a (morally and ethically) superior in-group that is perceived as legitimate and an out-group	*in-group/out-group* types [from a list of typical in-group and out- group types]
inferior out-group	superiority of in-group [Y/N]
dangerous out-group	perceived *threat* [Y/N]
hostile actions	- *hostility* (e.g., verbal attacks, belittlement, diminishment, discriminatory behavior) [Y/N] - instillment of *fear* of someone (out-group) [Y/N] - *incitement* to violence (against out-group) [Y/N]
not accepting any alternative views	intolerance [Y/N]
construal of Us/Them dichotomies	polarization/othering [Y/N]

Table 2: Annotation of extreme speech parameters

3.2. Other parameters of the annotation schema

3.2.1. The topics

For the purposes of the ARENAS project, the research focus was on content falling under the categories of Nation, Gender and Science: each case had to be classified according to one of the three categories, and any irrelevant cases would not be annotated further. The topics identified as belonging to these categories were the following:

> Nation**:** Nationalism, Neofascism/Neonazism, Economics, Globalization, EU, Ukraine-Russia War, Religion, Migration, Romaphobia, Islamophobia, Antisemitism
>
> Gender: Abortion, Reproductive Rights, Gender Equality, LGBTIQ+
>
> Science**:** Environmentalism, Vaccination, Nuclear Energy, Science Stance, Conspiracy Theories, Politically Motivated Revisionism

3.2.2. The poles of the polarized speech

It is crucial to determine both the speech author's in-group and the out-group s/he is targeting in order to classify extreme speech according to our working definition. Thus, 30 classifications were defined for the in-group/out-group categories within the ARENAS group. These included a combination of social, ethnic and/or religious groups that we had identified in previous research and through the literature review as likely targets of hate speech (e.g., Jewish community, LGBTQ+, etc.), additional labels for various types of organizations (e.g., NGOs, supranational actors), institutional roles (e.g., school staff, parents) and other groups that we hypothesized would emerge due to the specific scope of the ARENAS project (e.g., scientists/academics).

After the pilot study these categories could be expanded, reduced or amended according to feedback from annotators and the statistical results, while other labels emerging from the data could be added. This task was undertaken to ensure the most accurate and most relevant classification. A complete list from the pilot study is shown in Table 3.

Politicians/ Government	LGBTIQ+	Social Movements	Muslims	Parents
Extreme Right	Feminists	Media/ Journalists	Christians	Businesses
Right (politics)	The People	Scientists/ Academics	Other religious groups	Private Sector
Left (politics)	Nation/Own Country	Immigrants/ Asylum Seekers	Supranational actors (e.g., WHO, UN)	Public Sector
Extreme Left	Other Country	Jewish People	European actors	School Staff
Other Ideology	NGOs	Romani People	Family	Other/ Unclear

Table 3: In-group/Out-group categories

3.2.3. The tone of the speech

To facilitate ranking of the various degrees of offensiveness, we decided that – rather than attaching individual labels such as "offensive" or "abusive," and based on our previous work for another project (the IMsyPP project), we would label the 'tone of post' on a scale ranging from negative to positive (Baider 2023). This scale was adopted from the research by Poletto *et al.* (2019) since their study also considers the author's intentions in addition to the overall tone. This provides a more nuanced categorization of the tone of content, which was lacking in the negative-neutral-positive scale we used in previous annotations.

Label	**Meaning**
+1	Positive
0	Neutral, ambiguous or unclear
-1	Negative and polite, dialogue-oriented attitude
-2	Negative and insulting/abusive, aggressive attitude
-3	Strongly negative with overt incitement to hatred, violence or discrimination, attitude aimed at attacking or demeaning the target

Table 4: Tone and perspective of post scale

In IMsyPP, the different types of offensive speech were labelled as either *acceptable speech, speech offensive to the commentator,* or *speech offensive to a third party*. For the pilot study, we tested the extremism of posts with the categories developed by Udupa

and Pohjonen (2019), i.e., *Derogatory Extreme Speech, Exclusionary Extreme Speech* and *Dangerous Extreme Speech.* Our aim was to assess the necessity and the feasibility of specifying the degree of extremity. As the schema was to be used on datasets not limited to offensive or extreme speech, an option for non-extreme speech (which is broader than acceptable speech) was added to the original three-way classification task.

3.2.4. The narrative structure

To annotate the narrative structure, we scored the results using the units delineated in Table 5: *Character, Setting, Initiating Event, Internal Response, Plan, Attempt, Consequence, Helper, Opponent.* Each one was scored on a scale of 0-3, with 0 indicating an absence of the specific feature, and 3 an elaborate use of it. For example, 0 for *Character* would indicate that there is no naming of any agent, whereas 3 would indicate that there are two or more main characters in the narrative. Scoring relied on explicit lexical cues in the text in order to reduce disagreement among annotators. A detailed table regarding the scoring of the categories is available in Gillam *et al.* (2017). Our additions, *Helper* and *Opponent,* followed the same scoring rules as *Character.*

Category	**Definition**
Character	Agent(s) who performs an action.
Setting	Information about location or time (or setting the scene).
Initiating Event (IE)	Event(s) that motivate characters to take action (or causal event).
Internal Response	Feelings stated about the IE. They must be made by the character taking the actions related to the IE.
Plan	Thoughts stated by characters related to a decision to take action.
Attempt	Actions (to be) taken by characters motivated by IE.
Consequence	End result of characters' action in relation to the IE (consequential event).
Helper	Agent(s) who assist(s) the main characters on an action or plan.
Opponent	Agent(s) who challenge(s) or obstruct(s) the main characters regarding an action or a plan to be carried out.

Table 5: Proposed features for annotating the narrative structure

3.2.5. The emotions

As noted earlier, previous research has clarified the specific negative emotions that are associated with hate speech. These include emotions such as *Contempt* (Koselak 2005), *Disgust* (Baider 2019, 2022), *Fear* (Guillén-Nieto 2023) and *Anger* (El Sherief *et al.* 2018). Furthermore, in their research work to identify radical content online, Rehman *et al.* (2021) reported that negative emotionality (e.g., *Sadness* and *Jealousy*) can signal radical content. These emotions were added to our schema with a binary label (yes/no), in order to examine their prevalence in annotated content and their relationship to extremist content. *Pride* due to belonging to the in-group was also included as a positive emotion. The difference between the *internal response* in the narrative structure is that it had to be based on explicit cues, whereas the emotions mentioned above could be inferred.

3.2.6. The argumentation

In terms of rhetorical strategies, we were interested in detecting the use of *Appeal to Authority* as a way of enhancing the persuasiveness of the content; this tactic appeals to someone's alleged scientific, religious, or other expertise – a common strategy as observed in the IMsyPP dataset we had previously worked on. This was introduced in the schema with a binary label (yes/no).

To gain further insight into rhetorical modes, we used the IMsyPP classification of rhetorical means (Baider 2023) and used the following labels: *facts, examples, testimony, reasoning, conspiracy theories,* and *other.* It was decided then that a hierarchy would be provided in the annotators' guide book; this scale prioritized *conspiracy theories,* which were of most interest and relevance to extremist speech (Baider 2022, 2023), followed by *general reasoning*, and leaving the fine-tuned labels *facts, examples, testimony* for simpler content whose argumentation is not complex enough to fit the criteria for the *reasoning* category, and *other* for any cases that did not match any of the aforementioned labels.

In the last part of the schema, the multimodality of the content was annotated for the presence of *photos/images, video, links,* and *emojis*, with individual binary labels.

3.3. The annotation guide

The development of an elaborate annotation guide is fundamental to annotation work. It is also necessary for annotator training, as it ensures that annotators are clearly and regularly informed of the aims, definitions, and rules guiding the decision-

making process of every annotation task included in the schema. When there is a considerable degree of inter-annotator disagreement it can often signal a need to improve annotation guidelines.

We designed a guide that included relevant instructions, guiding principles and definitions for each task, e.g., what is an in-group and what is an out-group in relation to each of the emotions, rhetorical elements, types of extreme speech, etc. Furthermore, examples were provided to facilitate the annotator's recognition of relevant cases.

For broad categories such as *perceived threat,* more fine-tuned descriptions were provided, in this case elaborating on the various forms of threat, including physical (e.g., LGBTQI+ as sex offenders), economic (e.g., migrants as financial burden), health-related (e.g., vaccines cause autism), symbolic (e.g., migrants threatening cultural values).

For those tasks where annotators had to choose one label among several possible labels, explicit guidelines were given to help them determine the best option for ambiguous cases. For example, for *topic*, even if a case was relevant to both *Gender* and *Nation*, annotators were instructed to choose the label based on context and other cues that seem to have motivated the author to post. The annotation guide was revised for further clarifications after the pilot study and discussions with the annotators.

4. The pilot study

4.1. The process

The dataset of the pilot study consisted of 149 comments in Greek from the CONTACT online hate speech project (Assimakopoulos *et al.* 2017) and 101 comments in English from an earlier project (IMSyPP, Baider 2023). Comments were selected based on their previously identified topic, which was LGBTQI+, many of which included homophobic discourse. This boosted sampling approach was selected as it served to minimize irrelevant and/or non-offensive cases and to facilitate comparisons between the labels of earlier projects and the current study. Google spreadsheets were used as the annotation interface, with each line corresponding to one comment.

Each dataset was annotated by a team of four, two of whom had previous experience, while two were novices. All were local, and two belonged to the LGBTQI+ community. The rationale for these criteria was twofold. First, to see if there would be significant differences in inter-annotator agreement between the experienced and non-experienced annotators. Second, since comments were already identified as referring to LGBTQI+ issues, to explore if membership in an affected community would yield different results compared to

those not directly affected by the homophobic discourse, in a similar vein to the research of Maronikolakis *et al.* (2022). Individual training sessions were adjusted to each annotator's level of experience and examples were discussed.

Each comment was examined for all variables before the next comment was annotated. Where available, the metadata for each commenter, including username, was shared with the annotators; this provided context and facilitated the identification of individuals or group targets. Agreements between individual annotators were measured using pairwise correlations, followed by qualitative analysis of the results. Any observed discrepancies between the annotators indicated a need for more extensive training as well as more specific guidelines. Therefore, we organized a focus group in which selected annotators exchanged feedback; this helped us identify problems and develop more precise definitions. We will work to revise the schema and guidelines, after which we will hold a training program before proceeding to another round of annotations with partners from the ARENAS project. At this stage we will have important information related to intercultural and interlingual agreements and differences.

4.2. The results: Inter-annotator agreement

Once all annotators had completed their work, we examined the results and noted a number of discrepancies, concluding the following points:

- the emotions *sadness* and *jealousy* were difficult to identify;
- in the narrative structure, the labels *action, helper* and *plan* were not assigned consistently;
- in the extreme speech parameters schema, the labels *perceived threat, incitement to violence, superiority of in-group,* and *intolerance* were not consistently marked; moreover, we noted that the polarization label overlapped with *superiority of in-group*;
- no annotator used links and images.

These labels could therefore be better explained in our guidebook or not listed among emotions to be labelled since they may not be relevant to extremist speech.

Labels that were consistently identified included:

- the elements of the narrative structure such as *polarization/ othering, initiating event, in-group / out-group polarization, internal response*, *consequence, opponent*;
- the argumentative strategies of *appeal to authority, incitement to violence, intolerance*;
- the emotions of *anger, instillment of fear/ fear, disgust*, and *contempt,* which are also the emotions that have been identified as core to the emotion of hatred.

Given these insights, we plan to adjust the proposed schema by simplifying the choices and excluding the labels that were not identified (*jealousy, sadness*, *action, helper*, *plan*, *superiority of in-group*). Additionally, we will review the explanations for the labels *perceived threat* and *incitement to violence* since both are core to dangerous extreme speech/ hate speech and cannot be ignored. Last, as elaborated further in the next section, we suggest a differentiation between two types of speech: *directed* (*specific*) if it is addressed to a particular individual (insults such as *retard* and colloquial language); *generalised* if it is addressed most often to a specific ethnicity or (especially) religion (presence of numerals, lethal vocabulary).

4.3. Qualitative analysis

We will now offer some examples to highlight the importance of, as well as the difficulty inherent in labeling utterances. The quotes below were labeled acceptable or offensive in the IMsyPP project.

The quote "Y isn't there a straight pride march. Homophobia is a pathetic thing but having a parade to celebrate being gay is just as pathetic" triggered a series of diverse answers. After studying the annotations of these answers, we concluded that three changes to our labels might be possible.

Conclusion 1. Most quotes labeled offensive speech by a previous team were also labeled extreme speech by annotators, except when the extreme speech was addressed to an individual or toward a dominant community. The differentiation between generalized (i.e., addressed toward a community) or directed (i.e., addressed towards a specific individual) may be useful: thus, if insults and derision are directed at an individual or a dominant community, but there is no call for violence, the conclusion that the post is violent and contemptuous is still valid. We therefore decided that the label 'derogatory extreme speech' was unnecessary, because the label 'offensive speech' covers all cases.

Thus, when offensive speech is addressed to a vulnerable community, it is exclusionary (although it does not include a call for violence) extreme speech or hate speech (including a call of violence).

Examples:

Directed derogatory extreme speech toward an individual

(1) "The ONLY thing "pathetic" here is YOU. Sad and PATHETIC and really, really STUPID!"

This quote was not labeled extreme speech by one annotator because it is not directed to a specific community. Yet, after differentiating between *generalized* and *directed,* it can be considered extreme speech but "directed" to a specific individual.

Generalized derogatory extreme speech toward a dominant community, straight people

(2) "The reason there isn't a "straight pride" march is because GAY people are better than you straight people! Gays DESERVE a parade because of the way some of you A-hole straight losers treat them! (...)Think about THAT, you loser!"

The quote is both directed and generalized extreme speech towards straight people and one supposedly straight commentator. It is derogatory (inferiority – superiority of groups) and insulting (*assholes, losers*).

Conclusion 2. There seems to be no additional value in including the label *dangerous extreme speech*, since all quotes labeled dangerous extreme speech are actually overt or covert hate speech. These labels are much better known than dangerous extreme speech.

(3) "This is what the muslims believe. 'If a man has sexual relations with a man as one does with a woman, both of them have done what is detestable. They are to be put to death; their blood will be on their own heads'."

(4) "I can love them staying in their closet. I DON'T flaunt my sexual preference and that's exactly what they are doing today. YOU ARE THE ONE WHO MISSED THE BIBLE MESSAGES. QUOTE: A MAN WHO LIES WITH A MAN IS AN ABOMINATION"

The statements above were labeled dangerous speech by one annotator because: in comment 3 the death of the LGBTQ+ specific community is indirectly called for; in comment 4 the term 'abomination' could be labeled dangerous /hate speech since it encourages hatred and expresses the feeling of extreme disgust towards a community, both characteristics of hate speech.

However, another annotator labeled both quotes as derogatory speech or exclusionary speech, since they implied contempt and possibly exclusion based on protected characteristics. This annotator did not choose dangerous speech since there was no (i) direct call for harm, it (ii) did not have high degree of influence over audience. However, it has been noted that the semantic field of 'death' is found in hate speech. As for the degree of influence, comments on Facebook are widely disseminated.

Conclusion 3. As mentioned several times in this paper, examining comments out of their textual and social context easily leads to misinterpretation. For example, statement 5 below was labeled as exclusionary extreme speech by one annotator and acceptable speech by the other.

(5) "Happy people make me happy, all our love to the LGBT community! Thank goodness for normal people that have created this beautiful occasion!
NORMAL PEOPLE?
Yes normal. Those who accept and embrace our fellow humans in the LGBT community. Not discriminating on the grounds of gender, sexuality etc., etc.
That is exactly what I say Pattie. Normal people don't have to march since we know who we are and don't have to prove anything to anyone."

Conclusion 4. It is logical to think it is exclusionary if the quote refers to the LGBTQ+ community since the semantic field of abnormality is core to homophobia. However, according to the textual context it does not refer to homosexuality. The questioning of normality refers to people who do not accept different sexualities in the textual context.

These few examples underline the need to periodically discuss the labeling with the team, while they highlight the lengthy process required to establish a reliable and consistent annotation schema. Taking these steps will help to clarify different interpretations and will also reveal to the team why the presence of specific lexical fields should trigger a label of dangerous/ hate speech.

5. Final conclusion

This paper presents the results of a pilot study using a schema for annotating extreme narratives. First, we examined the common ground of both extreme speech and hate speech: both should be assessed within a spectrum of practices, and both contain a breach in the normative and civil behavior expected in a specific community. Hate speech, however, always includes a call for violence or hatred against a vulnerable community. Extreme speech can however express emotional grievances, feeling of victimization but also anger and fear towards *any* community. We also concluded that, in practice, the suggested new labels that distinguish among derogatory, exclusionary and dangerous extreme speech are not necessarily useful compared to other labels previously used in annotating hateful or offensive speech. This does not negate the importance of tagging and responding to extreme speech. Indeed, Bilewicz and Soral (2020) have revealed how derogatory language, which is part of extreme speech, leads to polarization and radicalization.

Psychological research has also shown the clear impact of contemptuous speech on activating already existing negative ethnic stereotypes (Kirkland *et al.* 1987) and amplifying these negative evaluations (Simon and Greenberg 1996) even if these sentiments were never publicly expressed. Extreme speech, which can be derogatory and uncivil therefore, offers a 'safe place' to spread racist views. However, in face of the desire to protect society harm, sanctioning extreme, uncivil or derogatory speech would not safeguard freedom of expression which includes freedom of expressing statements that offend, shock or disturb.[4]

References

Alkiviadou, N. (2019), "Hate speech on social media networks: towards a regulatory framework?", *Information & Communications Technology Law*, 28/1, p. 19-35.

Allport, G. W. (1954), *The nature of prejudice*, Addison-Wesley.

Ascone, L. and Longhi, J. (2018), "The expression of threat in jihadist propaganda", *Fragmentum,* 50, p. 85-98. https://doi.org/10.5902/2179219428823.

Assimakopoulos S., Baider F. and Millar S. (2017), *Online Hate Speech in the European Space*, Cham, Springer.

Baider, F. (2019), « Le discours de haine dissimulée; le mépris pour humilier », *Déviance et société*, 43/1, p. 71-100.

Baider, F. (2020), "Pragmatics lost? Overview, synthesis and proposition in defining online hate speech", *Pragmatics and Society*, 11/2, p. 196-218.

Baider, F. (2022), "Covert hate speech, conspiracy theory and anti-Semitism: linguistic analysis versus legal judgement", *International journal for the semiotics of law*, 35/6, p. 2347-2371, https://doi.org/10.1007/s11196-022-09882-w.

Baider, F. (2023), "Accountability Issues, Online Covert Hate Speech and the Efficacy of Counter-narratives", *Politics and Governance*, 11/2, p. 249-260.

Baider, F and Constantinou, M. (2020), "Covert hate speech: A contrastive study of Greek and Greek Cypriot online discussions with an emphasis on irony", *Journal of Language Aggression and Conflict*, 8/2, p. 262-287.

Baider, F. and Kopytowska, M. (2018), "Narrating hostility, challenging hostile narratives", *Lodz Paper in Pragmatics*, 14/ 1, p. 1-25.

Baider, F. and Romain, C. (2022), "Humorous remarks in covert hate speech and counter-speech", *Innovative Monitoring Systems and Prevention Policies of Online Hate Speech*, http://imsypp.ijs.si/wp-content/uploads/HumourBaider.pdf.

Bansal, V., Tyagi, M., Sharma, R., Gupta, V. and Xin, Q. (2022), "A Transformer Based Approach for Abuse Detection in Code Mixed Indic Languages", *ACM Trans. Asian Low-Resour. Lang. Inf. Process*, https://doi.org/10.1145/3571818

Bilewicz, M. and Soral, W. (2020), "Hate Speech Epidemic. The Dynamic Effects of Derogatory Language on Intergroup Relations and Political Radicalization", *Advances in Political Psychology*, 41/ 1, doi: 10.1111/pops.12670 0162-895X0.

[4] https://www.guidedroitshomme.fr/en/themes/hate-crimes-and-hate-speech/hate-speech/how-to-recognise-hate-speech

Brown, A. (2018), "What is so special about online (as compared to offline) hate speech?", *Ethnicities*, 18/3, p. 297-326. https://doi.org/10.1177/1468796817709846.

Carvalho, P., Cunha, B., Santos, R., Batista, F. and Ribeiro, R. (2022), "Hate speech dynamics against African descent People, Roma and LGBTQI communities in Portugal", in Calzolari, N. *et al.* (eds), *Proceedings of the Thirteenth Language Resources* (*LREC 2022*), European Language Resources Association, Marseille, p. 2362-2370.

Chakraborti, N. (2015), "Re-thinking Hate Crime: Fresh Challenges for Policy and Practice", *Journal of Interpersonal Violence*, 30/10, p. 1738-1754.

Council Framework Decision 2008/913/JHA of 28 November 2008. Entry into force 6 December 2008.

Das, A., Al Asif, A., Paul, A. and Hossain, M. (2021), "Bangla hate speech detection on social media using attention-based recurrent neural network", *Journal of Intelligent Systems*, 30/1, p. 578-591, https://doi.org/10.1515/jisys-2020-0060.

Davidson, T., Warmsley, D., Macy, M. and Weber, I. (2017), "Automated Hate Speech Detection and the Problem of Offensive Language", *Proceedings of the International AAAI Conference on Web and Social Media*, 11/1, p. 512-515, https://doi.org/10.1609/icwsm.v11i1.14955

Dixon L., Li, J., Sorensen, J., Thain, N. and Vasserman, L. (2018), "Measuring and mitigating unintended bias in text classification", Proceedings *of Conference on AI, Ethics, and Society*, p. 67-73.

El Sherief, M., Kulkarni, V., Nguyen, D., Yang Wang, W. and Belding, E. (2018), "Hate Lingo: A Target-Based Linguistic Analysis of Hate Speech in Social Media", *Proceedings of the International AAAI Conference on Web and Social Media*, 12/1, https://doi.org/10.1609/icwsm.v12i1.15041

El Sherief, M., Ziems, C., Muchlinsk, D., Anupindi, V., Seybolt, J., De Choudhury, M. and Yang, D. (2021), "Latent Hatred: A Benchmark for Understanding Implicit Hate Speech", in Moens, M.-F. *et al.* (eds), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Online, Association for Computational Linguistics, p. 345-363.

Fišer, D., Erjavec, T. and Ljubešić, N. (2017), "Legal framework, dataset and annotation schema for socially unacceptable online discourse practices in Slovene", in Wassem, A. *et al.* (eds), *Proceedings of the First Workshop on Abusive Language Online*, Vancouver, p. 46-51.

Fortuna, P. and Nunes, S. (2018), "A Survey on Automatic Detection of Hate Speech in Text", *ACM Computing Surveys (CSUR),* 51, p. 1-30.

Founta, A., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., Vakali, A., Sirivianos, M. and Kourtellis, N. (2018), "Large scale crowdsourcing and characterization of twitter abusive behavior", *Proceedings of the international AAAI conference on web and social media*, vol. 12/ 1.

Gagliardone, I., Patel, A. and Pohjonen, M. (2014), *Mapping and Analysing Hate Speech Online*, https://doi.org/10.2139/ssrn.2601792

Gillam, S. L., Gillam, R. B., Fargo, J. D., Olszewski, A. and Segura, H. (2017), "Monitoring indicators of scholarly language: A progress-monitoring instrument for measuring narrative discourse skills", *Communication Disorders Quarterly*, 38/2, p. 96-106.

Greimas, A. J. (1966), *Sémantique Structurale*, Larousse, Paris.

Greimas, A. J. and Courtes, J. (1979), *Semiotics and language: An analytical*

dictionary, Indiana University Press, Bloomington.
Guillén-Nieto, V. (2023), "The wording of hate speech prohibition: "You can't see the wood for the trees" ", in Guillen Nieto, V., Stein, D. and Doval Pais, A. (eds), *From Fear to Hate: Legal-Linguistic Perspectives on Migration*, Mouton de Gruyter, Berlin, p. 173-200, DOI: 10.1515/9783110789157-008.
Hymes, D. (1974), "Ways of speaking", *Explorations in the Ethnography of Speaking*, 1, p. 433-451.
Jones, S., Fox, C., Gillam, S. and Gillam, R. B. (2019), "An exploration of automated narrative analysis via machine learning", *Plos one*, 14/10, e0224634.
Kirkland, S. L., Greenberg, J. and Pyszczynski, T. (1987), "Further evidence of the deleterious effects of overheard derogatory ethnic labels: Derogation beyond the target", *Personality and Social Psychology Bulletin*, 13/ 2, p. 216-227. https://doi.org/10.1177/0146167287132007.
Knoblock, N. (2023), *Grammar of hate*, Cambridge Press, Cambridge.
Koselak, A. (2005), « Mépris/dédain, deux mots pour un même sentiment ? », *Lidil*, 32, p. 21-34.
Langton, R. (2018), "The Authority of Hate Speech", in Gardner, J., Green, L. and Leiter, B. (eds), *Oxford Studies in Philosophy of Law*, 3, Oxford University Press, Oxford, p. 123-152.
Machado Carneiro, B., Linardi, M., and Longhi, J. (2023), "Studying Socially Unacceptable Discourse Classification (SUD) through different eyes: "Are we on the same page?", https://arxiv.org/abs/2308.04180
Maronikolakis, A., Wisiorek, A., Nann, L., Jabbar, H., Udupa, S. and Schuetze, H. (2022), "Listening to Affected Communities to Define Extreme Speech: Dataset and Experiments?", in Muresan, S. *et al.* (eds), *Findings of the Association for Computational Linguistics: ACL 2022*, Association for Computational Linguistics, Dublin, p. 1089-1104.
Musolff, A. (2015), "Dehumanizing metaphors in UK immigrant debates in press and online media", *Journal of Language Aggression and Conflict*, 3/1, p. 41-56.
OHCHR (2013), *Rabat Plan of Action on the prohibition of advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence*, Morocco, 5 October 2012, UN Doc A/HRC/22/17 Add.4.
Omran, E., Al Tararwah E. and Al Qundus J. (2023), "A comparative analysis of machine learning algorithms for hate speech detection in social media", *Online Journal of Communication and Media Technologies*, 13/ 4, e202348.
O'Sullivan, P. B. and Flanagin, A. J. (2003), "Reconceptualizing 'flaming' and other problematic messages", *New Media & Society*, 5/1, p. 69-94. https://doi.org/10.1177/1461444803005001908.
Pohjonen, M. and Udupa, S. (2017), "Extreme Speech Online: An Anthropological Critique of Hate Speech Debates", *International Journal of Communication*, 11, p. 1173-1191.
Poletto, F., Basile, V., Bosco, C., Patti, V., Stranisci, M. (2019), "Annotating hate speech: Three schemes at comparison", *CEUR workshop proceedings*, 2481, p. 1-8.
Postmes, T., Spears, R. and Lea, M. (2000), "The Formation of Group Norms in Computer-mediated Communication", *Human Communication Research*, 26/3, p. 341-371.
Propp, V. J. (1928), *Morphology of the folktale* (2nd edition), University of Texas Press, Austin.

Rehman, Z. Ul, Abbas, S., Khan, M. A., Mustafa, G., Fayyaz, H., Hanif, M. and Saeed, M. A. (2021), "Understanding the Language of ISIS: An Empirical Approach to Detect Radical Content on Twitter Using Machine Learning", *Computers, Materials & Continua*, 66/2, p. 1075-1090.

Ricoeur, P. (1984), *Time and Narrative* (vol. 1), University of Chicago Press, Chicago.

Sap, M., Gabriel, S., Qin, L., Jurafsky, D., Smith, N.A. and Choi, Y. (2020), "Social Bias Frames: Reasoning about Social and Power Implications of Language", in Jurafsky, D. *et al.*, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, Association for Computational Linguistics, p. 5477-5490.

Simon, L. and Greenberg, J. (1996), "Further Progress in Understanding the Effects of Derogatory Ethnic Labels: The Role of Preexisting Attitudes Toward the Targeted Group", *Personality and Social Psychology Bulletin*, 22/12, p. 1195-1204, https://doi.org/10.1177/01461672962212001.

Stephan, W. G. and Stephan, C. W. (2000), "An integrated threat theory of prejudice", in Oskamp, S. (ed.), *Reducing prejudice and discrimination*, Lawrence Erlbaum Associates Publishers, p.23-45.

Timmermann, W. (2008), "Counteracting Hate Speech as a Way of Preventing Genocidal Violence", *Genocide Studies and Prevention: An International Journal*, 3/3, http://scholarcommons.usf.edu/gsp/vol3/iss3/8

Trifonas, P. P. (2015), "From Semantics to Narrative: The Semiotics of A. J. Greimas", in Trifonas, P. (ed.), *International Handbook of Semiotics*, Springer, Dordrecht, https://doi.org/10.1007/978-94-017-9404-6_50

Udupa, S., Gagliardone, I. and Hervik, P. (2021), *Digital hate: the global conjuncture of extreme speech*, Indiana University Press.

Udupa, S. and Pohjonen, M. (2019), "Extreme Speech and Global Digital Cultures", *International Journal of Communication*, 13, p. 3049-3067.

van Dijk, T. (1998), *Ideology: A Multidisciplinary Approach*, Sage, London.

Waldron, J. (2012), *The harm in hate speech*, Harvard University Press, https://doi.org/10.4159/harvard.9780674065086.

Walther, J. B., Anderson, J. F. and Park, D. W. (1994), "Interpersonal Effects in Computer-Mediated Interaction: A Meta-Analysis of Social and Antisocial Communication", *Communication Research*, 21/4, p.460-487, https://doi.org/10.1177/009365094021004002.

Wich, M., Al Kuwatly, H. and Groh, G. (2020), "Investigating Annotator Bias with a Graph-Based Approach", in Akiwowo, S. *et al.* (eds), *Proceedings of the Fourth Workshop on Online Abuse and Harms*, p. 191-199, Online, Association for Computational Linguistics.

Wodak, R. (2015), *The Politics of Fear: What right-wing populist discourses mean*, Sage Publications.

Wodak, R. and Richardson, J. (2022), "Anti-Sororism: Reviving the "Jewish world conspiracy" ", in Demata, M., Zorzi, V. and Zottola, A. (eds), *Conspiracy Theory Discourses*, John Benjamins, Amsterdam.

Yerden, A. U. and Turgut, K. (2024), "Using Artificial Intelligence Algorithms to Detect Hate Speech in Social Media Posts", *Ijonfest*, 2/1, p. 8-16, doi: 10.61150/ijonfest.2024020102.

Yuan, L. and Rizoiu, M-A. (2025), "Generalizing Hate Speech Detection Using Multi-Task Learning: A Case Study of Political Public Figures", *Computer Speech & Language*, 89, doi: https://doi.org/10.1016/j.csl.2024.101690.