JoSS
Journal of Speech Sciences

DOI: 10.20396/joss.v11i00.16153

THE TRUTH BELOW THE SURFACE: TOWARDS QUANTIFYING AND UNDERSTANDING THE EVALUATION OF GERMAN AND DANISH HATE SPEECH WITH EEG BIOSIGNALS

NIEBUHR, Oliver[1*]

NEITSCH, Jana[2]

[1]Centre for Industrial Electronics, University of Southern Denmark, Sønderborg, Denmark – ORCID: https://orcid.org/0000-0002-8623-1680
[2]Department of English Linguistics (IfLA), University of Stuttgart, Stuttgart, Germany – ORCID: https://orcid.org/0000-0002-2185-8829

Abstract: *The recipient is a stimulus-external factor that has so far hardly been investigated in hate-speech research. However, addressing this factor is essential to understand how and why hate speech unfolds its negative effects and which characteristics of the recipient influence these effects. The present study focuses on the recipient. Building on previous findings from explicit ratings and initial successful replications of such ratings through biosignals, we are conducting the first large-scale, systematic, and cross-linguistic biosignal study on hate speech based on two EEG measures: the beta-alpha ratio associated with arousal and the frontal alpha asymmetry associated with valence. A total of 50 Danish and German participants took part and were presented with spoken and written hate-speech stimuli, derived from authentic hate-speech posts on Twitter. Results show that Danes reacted more sensitively than Germans to hate speech containing figurative language (swear words), while Germans reacted more sensitively than Danes to hate speech with Holocaust references. In addition, teachers and lawyers showed less negative reactions to hate speech than church employees, students, and pensioners, which, despite small sample sizes, gives reason to think about whether social groups might react differently to hate speech. The effect of the presentation medium depended on the respective hate speech type. In particular, speaking out hate speech based on irony and indirectness attenuated its effects on recipients to such an extent that it is questionable whether the stimuli were still perceived as instances of hate speech at all. We discuss the results in terms of key tasks of future studies and practical implication for the punishment and management of hate speech on social media.*

Keywords: hate speech, EEG, biosignals, German, Danish, prosody, irony, Holocaust, swear words, foreigners, Muslims.

**Corresponding author: Oliver Niebuhr, University of Southern Denmark, Alison 2, DK-6400 Sonderborg, olni@sdu.dk*

1 Introduction

1.1 Hate speech research and the (neglected) role of the recipient

So-called bistable images have been dealt with in psychology for many decades (Long and Toppin, 1981; Gale and Findlay, 1983). Different types of these images are distinguished (see Rodríguez-Martínez and Castillo-Parra, 2018 for a summary), for example, those that are based on an inversion of the figure-ground relation (as in the case of the vase-face illusion, Fig. 1a) or those that invert a spatial or movement interpretation (as in the case of the Necker cube, Fig. 1b). Furthermore, there are also bistable images for which a change in the perceptual interpretation is associated with a change in the image's semantic content, as in the case of the rabbit-duck illusion in Figure 1(c).

Bistable images play a major role in psychology, because investigating who perceives which interpretation when and under what conditions provides indirect insights into how cognitive processes take place and how perceptual mechanisms work in the human brain (Rodríguez-Martínez and Castillo-Parra, 2018; Cao et al., 2018). Bistable images also say something about us as individuals. For example, viewers who perform better in creativity tests can often jump back and forth between the alternative interpretations of bistable images more quickly (Klintman, 1984; Laukkonen and Tangen, 2017). The same applies to younger as opposed to older people (Beer et al., 1989), females as opposed to males (Schechter et al., 1991), and bilingual as opposed to monolingual speakers (Bialystok and Shapero, 2005).

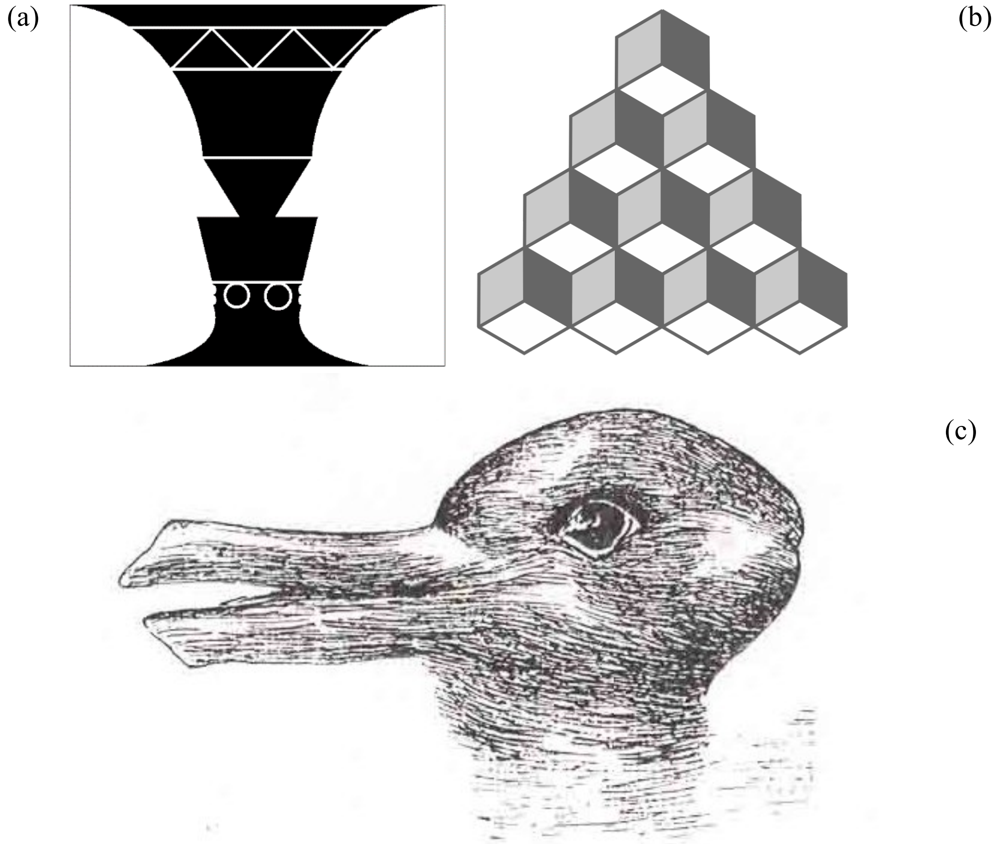

Figure 1: Three types of bistable images; (a) vase-face illusion, (b) Necker cube, and (c) rabbit-duck illusion. All images are taken from Wikimedia based on CC0 licenses.

Aside from the fact that the latter bilingualism effect is at least related to speech, what do bistable images have to do with hate speech? The answer is nothing at all – in strict phenomenological terms. But the comparison of hate speech to bistable images is still useful insofar as it helps us face two facts: First, perception is always something subjective, something constructed that is shaped by knowledge, experience, and situational contexts, see the numerous examples across modalities in Goldstein (2008), Bregman (1994), and Handel (1989). Second, this first fact is currently hardly reflected in the societal, political, and scientific discussion of hate speech. It is of course important that there are general definitions of hate speech, such as that of the United Nations (see Peters, 2020), on which the present chapter is based as well[1]. At first glance, this definition appears objective and concrete. But, it actually only states which types of perceptual interpretations are to be classified as hate speech – provided that certain linguistic and/or semantic criteria are met as well. Hate-speech definitions of social-media companies such as Facebook and Twitter are no better in this regard, see MacAvaney et al. (2019). No definition specifies how these interpretations arise and what triggers them. In terms the rabbit-duck illusion, the existing definitions of hate speech would roughly state: the duck with the aggressively open beak is hate speech, but the rabbit with the big set-back ears is not. Or, in terms of the hate speech definition: If you feel attacked and/or if you consider a certain wording pejorative, then the corresponding stimulus is an example of hate speech; in all other cases it is not.

But how does the interpretation of either duck or rabbit come about? Or, In terms of the hate speech definition: what causes the perception of being attacked or of being slighted by pejorative language? Current research and development approaches try to derive this interpretation from the stimulus itself. Malmasi and Zampieri (2017), for example, use large, annotated corpora to facilitate automatic hate-speech identification with reference to n-gram-based lexical baselines. Gambäck and Sikdar (2017) follow a similar machine-learning, big-data approach and find that word vectors formed on a semantic basis provide a better hate-speech identification performance (relative to the gold standard of human ratings) than randomly compiled word vectors or n-grams, which are generated without reference to semantics on the basis of individual letters, see also Waseem and Hovy (2016), Davidson et al. (2017) or Ruwandika and Weerasinghe (2018). The study by Martins et al. (2018) is more emotion-oriented and uses a Natural Language Processing (NLP) approach. Ultimately, however, it is also based on big data and a stimulus-related identification strategy. Fortuna and Nunes (2018), MacAnaney et al. (2019) and, more recently, also Papcunová et al. (2021) give excellent overviews of the status of this type of hate speech research.

Classical linguistic approaches to the identification of hate speech are comparatively rare or, at least, not very visible, which may be due to the inherently digital nature of social media posts or the fact that the symbols and morphosyntax in these posts are somewhat inaccessible to linguists as they often differ greatly from standard written language. However, there are linguistic studies on hate speech. Balcerzak and Jaworski (2015), for example, determined various adjectives, adverbs, and proper names that characterize hate speech posts in American English. The analysis by Jaki and De Smedt (2019) of German right-wing posts on Twitter provides still more concrete insights into which wordings and parts-of-speech are typical of hate speech. Geyer (2019, 2021) carried out linguistic analyses for Danish on the use of metaphors in hate speech and on grammatical patterns that characterize hate speech, see also Bick et al. (2017). Calderón et al. (2021) points out a problem in this context that also applies to machine-learning approaches: Hate speech changes quickly. Authors of hate-speech posts are inventive in developing new word forms and clauses, as well as new forms of expression in general,

[1] The UN defines hate speech as “any kind of communication in speech, writing or behaviour, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, colour, descent, gender or other identity factor.” (https://www.unhcr.org/5df9f0417.pdf)

which are intended to circumvent the hate-speech search and identification efforts of social-media companies and public authorities.

Regardless of this problem, we consider the stimulus-oriented approach to be insufficient. It remains at the level of the sender or the signal. The real problem does not arise there, but at the level of the recipient. Put simply: hate speech is whatever the recipient perceives as hate speech. Of course, stimuli contain triggers of or, rather, cues to hate-speech interpretations that need to be detected and understood. But there is also the individual recipient with his/her specific emotions, experience, knowledge, situational context, etc. Thus, coming back once more to the rabbit-duck illusion, it is not enough to try to determine whether the circular structure in the center of Figure 1 (c) is an eye, and if so, to determine subsequently with reference to adjacent (contextual) features whether it is a left eye (duck) or a right eye (rabbit). We also need to turn to the recipient and understand the conditions under which s/he interprets the circular structure as a left or a right eye. Only in this way can general hate-speech definitions like "any kind of communication that attacks ..." (see footnote 1, cf. also MacAvaney et al., 2019) be further substantiated and connected to appropriate, recipient-specific (legal) consequences for the originators of hate-speech stimuli.

1.2 Own prior work

XPEROHS is an international, German-Danish research project that examines hate speech from a cross-linguistic point of view and, most importantly, represents perhaps the first project in which the links between linguistic features and their impact on recipients are systematically investigated in perception experiments. The aim of XPEROHS is to provide social and political decision-makers with more concrete guidelines on what hate speech really is (Baumgarten et al., 2019).

For the perception experiments, a set of several million authentic Twitter and Facebook posts was compiled and turned into a tagged, author-anonymized corpus (Bick, 2020). This corpus served to determine, amongst others, which types of hate speech expressions are particularly common in German and Danish. The same six types of hate speech emerged in both languages: irony (IRO), rhetorical questions (RQ), imperatives (IMP), figurative language (FGL) in the form of swear words, Holocaust references (HOL) and indirectness (IND), which is based on introductory phrases like "I have nothing against ___, but___".

For our own prior work, a sample of 12 stimuli was selected from the XPEROHS corpus for both languages. The stimuli belonged to none of the above types, but were classified in separate pre-tests per language as clear instances of hate speech by both linguistics and naive readers. The stimuli were moreover selected to address the two major target groups of hate speech in German and Danish: foreigners in general and Muslims in particular. Six of the 12 stimuli in the selected sample of each language targeted foreigners, the other six Muslims, see Neitsch et al. (2021). The 12 stimuli are henceforth referred to as ORIG stimuli.

After the ORIG stimuli had been defined, they were used in the next step of the stimulus generation to create all six of the above-mentioned hate-speech types. For each type, this derivation process used a constant morphosyntactic strategy, such as attaching a preceding or following trigger phrase to the ORIG stimuli. The process resulted in a set of 12 x 7 or 84 stimuli per language; 12 stimuli each for IRO, RQ, IMP, FGL, HOL, and IND – plus the 12 ORIG stimuli.

The present study is also based on these 84 German and Danish stimuli, see method section 2.1. The stimuli were presented to participants in written form as well as in spoken form. The latter stimuli were created by eliciting them in an informal speaking style by one experienced speaker per language who, moreover, represented the typical originator of hate speech in Western culture (male, Caucasian, 30-45 years old, cf. Hrdina, 2016). Informal here means that the hate speech stimuli were not realized in a shouting, angry or generally highly expressive way, but rather casually, as would be the case in a conversation between two familiar speakers in a pub, for example. For further details on the prosody of the hate-speech stimuli, see Neitsch and Niebuhr (2020).

Neitsch and Niebuhr (2019) developed a 2D Rating Space for their perception experiments so that participants were able to rate each stimulus with a single mouse click along two different criteria: their own attitude towards the stimulus (x axis, “degree of unacceptability”) and what the appropriate societal reaction to the stimulus should be (y axis, "strength of the consequences for the originator"). The 2D rating space is shown in Figures 2(a)-(b), together with obtained key results. In general, the insights gained from our own prior work on hate-speech perception can be summarized as follows.

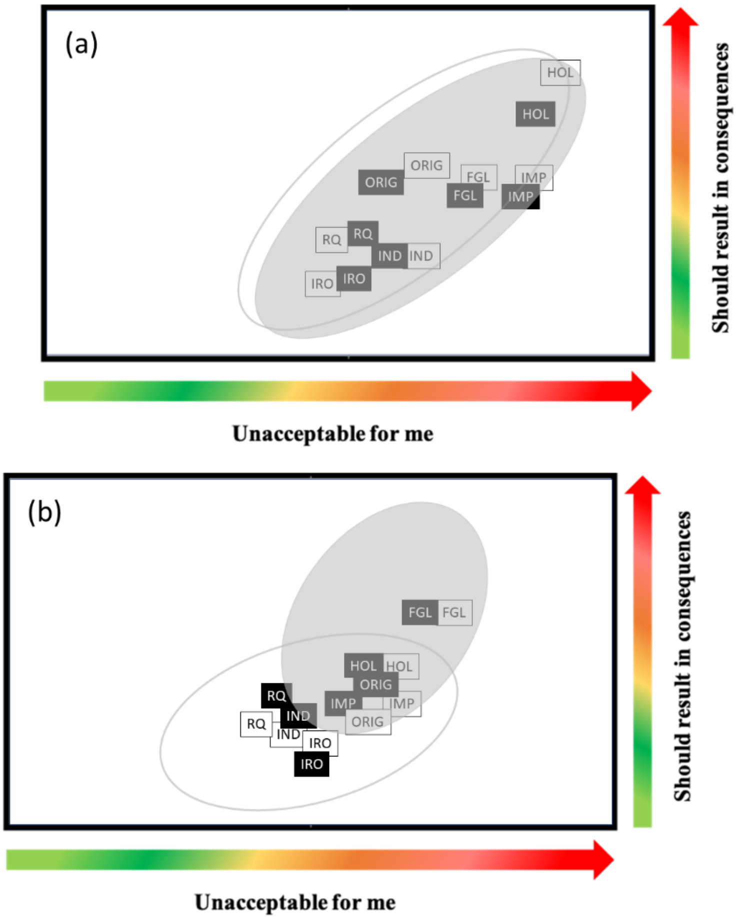

Figure 2: Mean values of participant’s stimulus ratings (N=28) along the two axis per stimulus type. White boxes refer to spoken, black boxes to written hate-speech stimuli. The filled and unfilled grey circles refer to the range of ratings for foreigner- and Muslim-directed hate speech, respectively. The top panel (a) shows the results for German, the bottom panel (b) those for Danish.

First, hate speech is not a homogeneous phenomenon that is categorically distinct from non-hate speech. Rather, a continuum of perceived severity emerged across the hate-speech stimuli in both German and Danish.

Second, the stimulus medium makes a significant difference in both languages. It also interacts with the type of hate speech. In short, if stimuli contained a strong lexical trigger for hate speech – for example, a call to action like "throw them out" (IMP) or wordings like "into the KZ" (HOL) or "Muslim crap" (FGL) – then the perceived severity of spoken hate speech was greater than that of written hate speech. In other words, written hate speech with strong lexical triggers becomes worse when spoken out loud. By contrast, when it comes to the IRO and RQ types of hate speech that are mainly expressed through prosody, then the spoken form is the one that is significantly less severe than the written form.

Third, the recipient matters as well. Unlike the Danes, for example, the Germans reacted very strongly to hate speech with Holocaust references (HOL). The Danes, on the other hand, perceived hate speech with swear words (FGL) to be significantly worse than the Germans. In addition, the Danes differentiated more between foreigner- and Muslim-directed hate speech. The latter was rated as significantly worse, especially along the y-axis, which demands consequences for the originator. The Germans, by comparison, hardly differentiated between foreigner- and Muslim-directed hate speech, and if they did, then rather along the x-axis of personal unacceptability – but, like, the Danes, also to the detriment of Muslim-directed hate speech.

These country-specific reactions are presumably a result of culture and social imprinting. For instance, because of their WWII history, the Germans learn a lot more than the Danes about the horrors of the Holocaust in schools and the media. That country of origin significantly influences the perception of roughly identical stimuli is a first example of why it is important to better understand the recipient side in the definition, identification and evaluation of hate speech. Recently, Neitsch and Niebuhr (2023) demonstrated another influencing factor on the recipient side by showing – in line with other studies on fear and crisis management – that hate speech is perceived as less severe if participants are not expose to it alone, but in the presence of a good friend.

Fourth, Neitsch and Niebuhr (2020, 2021) carried out pilot studies with biosignals and found a high level of correspondence between the explicit evaluation of hate speech stimuli in the 2D rating space on the one hand and the participants' direct but implicit physiological reaction to the stimuli on the other. Neitsch and Niebuhr concluded that biosignals can be used instead of explicit ratings to study the evaluation of hate-speech stimuli. Biosignals probably should even be used, as they offer several advantages over explicit ratings: (i) The participants do not have to perform any unnatural meta-linguistic tasks. Instead they just read the stimuli or listen to them passively like in an everyday situation. (ii) Biosignals can provide more detailed insights into hate speech evaluations; they allow to make parametric, physical measurements of several signals simultaneously and, if useful, analyze their time course relative to the stimulus. (iii) Bio-signals show the recipient's honest reaction or evaluation; i.e. they avoid a bias due to participants giving "socially desired responses", which is to some degree inevitable in any explicit rating task, even in anonymous ones (see also Neitsch and Niebuhr 2023).

In short, compared to explicit ratings, biosignals increase both the internal and the external validity of the obtained results of a hate-speech perception experiment. Neitsch and Niebuhr (2020, 2021) tested 5 different types of biosignals: EEG, heart rate, skin-conductance response (SCR), respiration (RespTrack) and pupillometry. Heart-rate measurements were found to be rather insensitive, and pupillometry was difficult to apply in comparable ways to both written and spoken hate speech. SCR measurements did not have any of these disadvantages. However, they required participants to sit unnaturally still in a chair, because movements (and the drying airstream they create at the sensor) strongly influenced SCR measurements. Niebuhr and Neitsch therefore recommended using EEG and RespTrack measurements to analyze people's reactions to hate-speech stimuli.

1.3 Questions

Against the outlined background, the aim of the present study is to replicate and refine the hate-speech findings for German and Danish summarized in 1.2 and in Figure 2(a)-(b) by means of biosignals. To the best of our knowledge, this is the very first study that investigates hate-speech perception and evaluation with biosignals on a large-scale, systematic, and cross-linguistic basis. Although we collected both EGG and respiratory data in the experiment reported below, we focus here on the EEG signals for two reasons: Firstly, the data from the two biosignal sources provide largely converging results and, secondly, the connection between EEG and emotional reactions is better understood and more established (cf. the references in 2.4 related to the MUSE II measures).

Our research questions are as follows:

- (I)Do the biosignals obtained mirror our earlier explicit-rating results by also forming a reaction continuum for the evaluation of the hate speech stimuli, for example, in terms of differences in the levels or amplitudes of EEG measurements?
- (II) With regard to (I), do we find the same German-Danish differences in the evaluation of hate speech with regard to the target group (foreigners vs. Muslims) and the type (e.g., HOL vs. FGL)?
- (III) With regard to (I), do we find supporting evidence for the relevance of the hate-speech medium, for example, in terms of some hate-speech types causing stronger physiological reactions in the spoken domain (e.g., HOL, IMP, FGL) and other types in the written domain (e.g., IRO, RQ)?
- (IV) Based on (II)-(III), do we find indications that prosody is able to attenuate hate speech to such an extent that it no longer has to be classified as hate speech?
- (V) Beyond the replication of previous findings, what additional insights do the EEG biosignals provide – for example, with regard to “socially desired responses” and inter-individual variation?

2 Method

2.1 Stimulus material

The selection of the 12 ORIG (base) stimuli from the authentic posts in the XPEROHS corpus (Bick et al. 2020) and the concept of deriving the six most frequent German and Danish hate speech types from each of these ORIG stimuli (cf. Neitsch and Niebuhr 2021) has already been described in 1.2. Therefore, Table 1 below only shows two examples of the derivation of the IRO, RQ, IMP, FGL, HOL, and IND stimuli from the ORIG stimuli as a supplement to the explanations in 1.2. One example concerns Danish (based on the target group of Muslims), the other one German (based on the target group of foreigners).

Table 1: Examples of German and Danish ORIG stimuli and the six further types of hate speech derived from them according to constant, type-specific concepts.

Type	Concept	German	Danish
ORIG	Selected authentic baseline post – no changes made.	*Diese Ausländer bringen doch nur ihre Kriege hier in unser Land!* (These foreigners only bring their own wars here to our country!)	*Den eneste integration muslimer ønsker, er den i vores velfærdssystem!* (The only integration Muslims want is that into our welfare system!)
IRO	The original meaning was turned into irony by adding adverbs	*Die ach so friedvollen Ausländer würden ihre Kriege ja niieee bei uns austragen!*	Muslimer kunne da aaaldrig finde på kun at ønske integration i vedfærdsydelserne!

	like 'never', often exaggerated by repeating letters ('neeever') and/or modal particles or associated constructions like 'the oh so…'	(The oh so peaceful foreigners would neeeever fight their own wars here in our country!)	(Muslims would neeever consider to integrate themselves only into our welfare system!)
RQ	A rhetorical question was always added at the end of the original stimulus.	*Diese Ausländer bringen doch nur ihre Kriege hier in unser Land! Wer will denn hier Krieg haben?* (These foreigners only bring their own wars here to our country! Who wants to have a war here?)	*Den eneste integration muslimer ønsker, er den i vores velfærdssystem!* Hvem vil ikke gerne være en del af vores velfærds-system? (The only integration Muslims want is that into our welfare system! Who would not like to become part of our welfare system?)
IMP	A separate imperative sentence was added either before or after the ORIG stimuli (mostly after it). In some cases the syntax of the ORIG stimulus was changed to that end.	*Schmeißt die Ausländer raus aus Deutschland! Sie bringen nur ihre Kriege hier in unser Land!* (Expel all foreigners from Germany! They only bring their own wars here into our country!)	*Giv ikke muslimer adgang til vores velfærdssystem! Det er jo det eneste, de vil integreres i!* (Do not give Muslims access our welfare system! After all, that's the only thing they want to be integrated into!)
FGL	The words for 'foreigner' or 'Muslim' were either supplemented by a slur like 'scum' or 'rat' or, where it fitted the context better, replaced by a new term that previous studies found in hate-speech posts.	*Dieser Ausländerdreck bringt doch nur seine Kriege hier in unser Land!* (This foreigner scum only brings its own wars here to our country!)	*Den eneste integration perkere ønsker, er den i vores velfærdssystem!* (The only integration 'perker' [slur word for Muslims] want is that into our welfare system!)
HOL	Holocaust references were mostly made by adding a separate (elliptic) sentence after the ORIG stimulus, demanding to send the respective target group into a concentration camp.	*Diese Ausländer bringen doch nur ihre Kriege hier in unser Land! Ab ins KZ mit ihnen!* (These foreigners only bring their own wars here to our country! [Throw them] all into a concentration camp!)	*Den eneste integration muslimer ønsker, er den i vores velfærdssystem! Send dem alle til et kz-lejr!* (The only integration Muslims want is that into our welfare system! [Throw them] all into a concentration camp!)
IND	The ORIG stimulus was introduced by a sentence such as 'I have nothing against Muslims /foreigners, but...'	*Ich hab ja nichts gegen Ausländer, aber die bringen doch alle nur ihre Kriege hier in unser Land!* (I have nothing against foreign-ers, but they all only bring their own wars here to our country!)	*Jeg har ikke noget imod muslimer, men det eneste, de ønsker, er deres integration i vores velfærdssystem!* (I have nothing against Muslims, but the only thing they want is their integration into our welfare system!)

2.2 Participants

A total of 50 people took part in the experiment, 25 Danes and 25 Germans. Both samples consisted of different subgroups in order to cover as many parts of the population as possible. Specifically, the German and Danish samples both included five pensioners, five lawyers, five employees of church institutions (e.g., pastors, organists, deacons), five teachers (or lecturers), and five students. Table 2 summarizes the key data of the subgroups in the two samples.

Table 2: Key data of the five difference subgroups (of five persons each) in the Danish and German participant samples; y=yes, n=no.

Sample	Subgroup	ex (m/f)	Average age	Familiar with term 'hate speech' (y/n)	Experience with hate speech (y/n, if y: $1^{st}/2^{nd}$ hand)
German	pensioners	m, 2f	74.0 years	2y, 3n	0y, 5n
	lawyers	m, 1f	41.5 years	5y	1y (2^{nd} hand), 4n
	church employees	m, 2f	55.6 years	4y, 1n	0y, 5n
	teachers	m, 3f	36.6 years	5y	2y (2^{nd} hand), 3n
	students	m, 3f	18.9 years	5y	3y (1^{st} and 2^{nd} hand), 2n
Danish	pensioners	m, 3f	67.8 years	4y, 1n	0y, 5n
	lawyers	m, 2f	44.9 years	5y	2y (2^{nd} hand), 3n
	church employees	m, 2f	38.4 years	4y, 1n	2y (2^{nd} hand), 3n
	teachers	m, 4f	29.3 years	5y	3y (1^{st} and 2^{nd} hand), 2n
	students	m, 3f	19.2 years	5y	4y (2^{nd} hand), 1n

Table 2 shows that the German and Danish samples are overall similar in terms of the proportion of male and female participants, their average age range, and their familiarity with the term 'hate speech'. The majority of participants in both samples knows the term (> 80 %). However, while the majority of German participants stated to not have any experience with hate speech (either 1st or 2nd hand), almost half of the Danish participants (44 %) have experienced hate speech before, although mostly 2nd hand. The German and Danish students represented the only two subgroups who were 100 % familiar with the term 'hate speech', and a majority of them had moreover already experienced hate speech 1st hand (i.e. they were targets themselves) or 2nd hand (i.e. they experienced others being a target). The only other comparable subgroup were the teachers. The lawyers and church employees mostly stated to know the term, but to have no direct or indirect experience with it. Only the pensioners had never experienced any hate speech and a lot of them (especially in the German sample) have also not heard of the term 'hate speech' before.

2.3 EEG system

Participants' perceptual interpretation of hate speech was measured by means of Electroencephalography (EEG). The EEG measurements were taken using the MUSE II headset, which was connected to the Muse Monitor app via Bluetooth (Richer et al., 2018). As Figure 3(a) shows, the MUSE II is a kind of headband in which four dry electrodes are embedded, two each on the left and right side of the frontal lobe and the temporal lobe (Fig. 3b). According to the standard reference system for placing EEG electrodes over the brain (Klem et al., 1999), the MUSE II headset measures the brain activity at AF7, AF8 (Fig. 3c) as well as at TP9 and TP10.

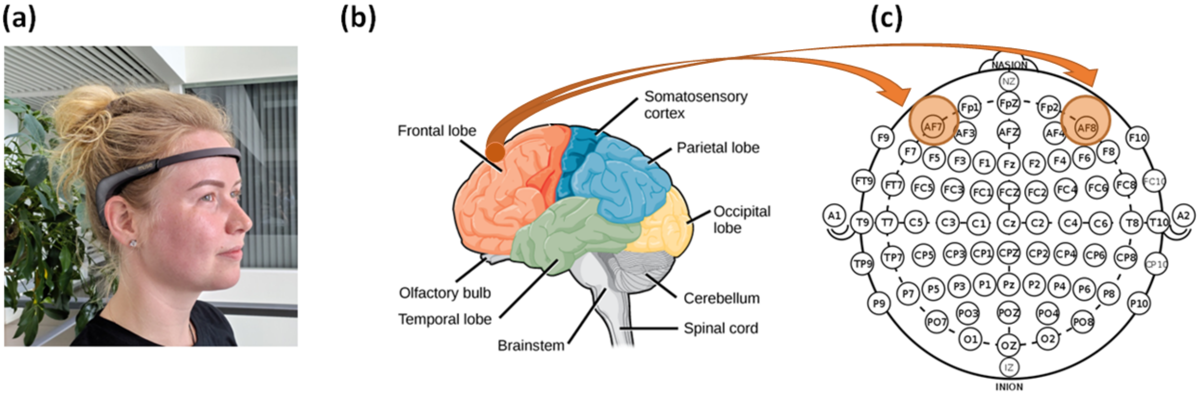

Figure 3: (a) Photo of a test participant wearing the MUSE II EEG headset; (b) Sagittal view of the brain showing the positions of frontal lobe (orange) and temporal lobe (green); (c) Position of the relevant measuring electrodes AF7 and AF8 in the frontal lobe. Figure parts (b) and (c) have been modified under CC licenses from Wikimedia (CNX OpenStax 2016; Oxley 2017).

The Muse Monitor software conducts a spectral analysis of the raw EEG signals at each of the four electrodes. The signals are broken down by the spectral analysis into five frequency bands: Delta (<4 Hz), Theta (4-7 Hz), Alpha (8-15 Hz), Beta (16-31 Hz), and Gamma (> 31 Hz), see Garcia-Moreno et al. (2020). For each frequency band the logarithm of the power spectral density is then calculated in dBµV at a sampling rate of 256 Hz. The dBµV values can vary between -1 and +1 and are used as raw measurements in the present study. A meta-study by LaRocco et al. (2020) shows that the MUSE II headset provided reliable and precise measurements over a large number of studies. That is, in all studies the measurements turned out to be highly correlated with participants' perceptual impressions, which were independently determined by means of behavior changes or rating tasks (Asif et al., 2019; Garcia- Moreno et al., 2020; Herman et al., 2021).

2.4 EEG measurements

For the purpose of the current experiment, the data collection was limited to the two frontal lobe electrodes AF7 and AF8 (see Fig. 3b-c), since this area of the brain is (more than the temporal lobe) associated with attention, language, speech, emotion, personality, and moral as well as social reasoning (Chayer and Freedman 2001) – all of which are properties and processes potentially relevant in the perceptual evaluation of hate-speech stimuli.

Our data collection was furthermore based on the assumption that the perceptual evaluation of hate speech is reflected in the way participants experience stress and emotions. For the choice of measurements, this assumption meant, firstly, that we focused on measuring two different frequency bands at AF7 and AF8, the alpha band and the beta band. The frequency energy in these two bands is most closely associated with stress and emotions, particularly negative ones (Herman et al., 2021; García-Acosta et al., 2021; Zhang et al., 2018; Zhao et al., 2018).

Secondly, the way in which participants experience stress and emotions can be operationalized as consisting of (at least) two different dimensions: the degree of arousal, and the degree to which this arousal is positive or negative. The latter dimension is also known as valence (García-Acosta et al., 2021). The two dimensions, arousal and valence, can be mapped onto separate measures that are derivable from the raw dBµV data taken at AF7 and AF8: Following the results of previous studies, we measured arousal in the form of the beta frequency energy in relation to the alpha frequency energy, also known as the the beta-alpha ratio or BAR. The energies at AF7 and AF8 were added up to that end at each point in time. The higher this

BAR value, the more was the participant aroused by the stimulus (García-Acosta et al., 2021). The effect of a stimulus on valence was determined in terms of the frontal alpha asymmetry or FAA (Zhao et al., 2018; García-Acosta et al., 2021; Zhang et al., 2018). FAA represents the energy ratio between the left frontal-lobe electrode AF7 and the right frontal-lobe electrode AF8 in the alpha frequency band (AF7/AF8). The lower this FAA value is below 1, the more negative was the valence triggered by the stimulus.

2.5 Experimental procedure

Before the actual experiment, the 25 participants in the German and Danish sample were randomly assigned to one of two experimental lists. That is, 50 % of the participants began with the spoken stimuli and then proceeded to the written stimuli. The other 50 % of participants started with the written stimuli, followed by the spoken ones. To avoid artifacts due to fatigue or habituation, there were several days (usually about a week) between the two lists.

The experiment was mainly carried out in the acoustics laboratory of the Center for Industrial Electronics (CIE) at the University of Southern Denmark[2]. Participants took part in the experiment in individual sessions. In each session, the participant sat on a comfortable office chair in a sound-attenuated environment in front of a PC screen, see Figure 4. The experiment began with a warning about the potentially disturbing nature of the stimuli. The participant had to click a button on the screen to acknowledge that s/he had read this warning. At the same time, s/he gave the informed consent to participate with this mouse click. Then, s/he was forwarded to an input mask through which general personal data was anonymously requested for statistical analysis purposes (see Table 2). The following information screen informed the participant about the task. It was stated that the task would simply be to expose oneself in a focused but authentic and natural way to the stimuli, which would be presented in either written or spoken form.

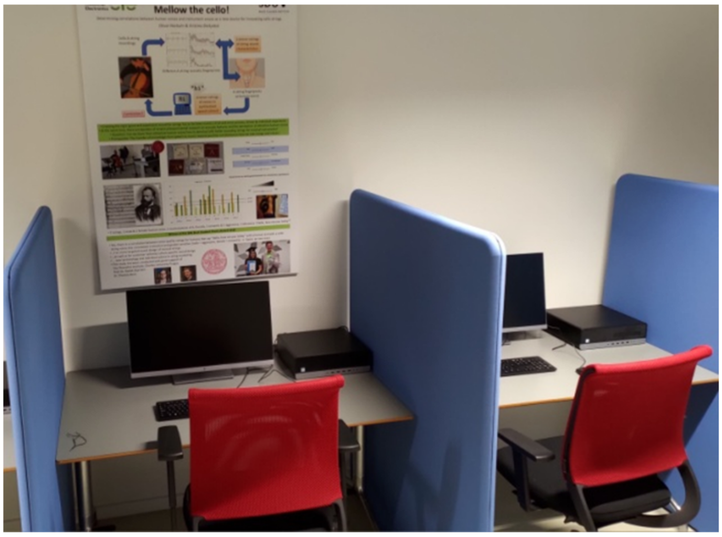

Figure 4: Sound-attenuated workplace at the CIE acoustics laboratory, where the experiment was carried out with participants in individual sessions of about 20 minutes (cf. footnote 2).

In the next step, the participant was asked to switch on and put on the MUSE II headset as well as the headphones (Quite Comfort 35 II), both of which were provided on the table in front

[2] Due to pandemic restrictions and measures, some participants also carried out the experiment at the first author's home in the office, but under similar acoustic conditions and with the identical equipment.

of the participant. The active noise cancellation of the headphones was switched off. The headphones were used in the condition of the spoken stimuli, but were also worn in the condition of the written stimuli in order to avoid that the special acoustic effect of wearing headphones could become a confounding factor in the experiment.

The 12 stimuli of each type were presented in blocks, i.e., for example, first the 12 FGL stimuli, then the 12 RQ stimuli, etc. Within each block, the stimuli targeting foreigners and Muslims were also presented as coherent sequences. The order of presentation was randomized across the participants, both at the block level and at the sequence level. The purpose of this block-/sequence-wise presentation mode was to continuously expose the participant to a single stimulus type for about 30 seconds. This time interval was (e.g., in Neitsch and Niebuhr, 2020) found to be long enough for the corresponding stimulus condition to shape the EEG measurements according to the participant’s hate-speech evaluation.

Prior to presenting the first stimulus block, the participant was additionally instructed to sit still and do nothing for 30 seconds. This rest phase served as a zero-stimulus reference condition (henceforth REF) to which the measurements of all stimulus conditions could be compared. It was moreover important to check for differences between the REF measurements of the two stimulus-medium conditions 'written' and 'spoken'. Only if the two REF conditions provided identical values could measurement differences be reliably interpreted as effects of the stimulus medium and not as offset differences between two separate experimental sessions.

The experimenter left the room before the zero-stimulus reference condition (REF) and only came back in after the last stimulus block. A complete experiment session lasted about 20 minutes, including the participant’s briefing and de-briefing.

2.6 Variables and statistical analysis

The experimental design included 2 dependent variables: the beta-alpha energy ratio BAR (integrating the AF7 and AF8 data), with higher values indicating a higher level of arousal in response to the stimuli, and the frontal alpha asymmetry FAA whose values can increase above 1 or decrease below 1, in this way indicating a stronger positive or negative valence, respectively.

We investigated how these two dependent variables are affected by four independent variables: Type, i.e. the seven hate-speech conditions ORIG, FGL, IRO, RQ, IMP, HOL, IND plus the zero-stimulus reference condition REF; Target, i.e. foreigners in general vs. Muslims in particular; Medium, i.e. the spoken vs. the written hate-speech stimuli; and Sample, i.e. German vs. Danish.

The statistics applied to this experimental design was a multivariate general linear model. It consisted of the three within-subjects factors Type, Target, and Medium as well as of the between-subjects factor Sample. Furthermore, we included Group, i.e. the different kinds of participants summarized in Table 2, as a covariate in the statistical model. In the results section below, we report within-subjects statistics with Greenhouse-Geisser corrections, if required. Multiple-comparisons tests between factor levels are reported with Sidak corrections included. Main effects and interaction statistics are reported based on the Wilks' lambda test statistic, following Ateş et al. (2019). The corresponding effect-sizes are reported in terms of partial eta-squared (η_p^2), i.e. proportion of residual variance attributable to effect or interaction after other factors are partialized out from the total non-error variation. Partial eta-squared is a suitable effect-size measure for our statistical model (Dattalo, 2013:33) and in fact the derfault measure in SPSS. We conducted our statistical tests by means of SPSS v. 28.0

3 Results

The results of the multivariate general linear model showed significant main effects of all fixed factors but Medium. In terms of effect sizes, the strongest main effect was that of Type ($F[14,34] = 112.432$, $p < 0.001$, $\eta_p^2 = 0.979$) followed by Target ($F[2,46] = 176.665$, $p < 0.001$, $\eta_p^2 = 0.885$) and the covariate Group ($F[2,46] = 35.378$, $p < 0.001$, $\eta_p^2 = 0.606$). The overall weakest main effect was that of Sample ($F[2,46] = 4.683$, $p = 0.014$, $\eta_p^2 = 0.169$). Table 3 summarizes all relevant two-way and three-way interactions associated with these main effects. As can be seen, Medium was also not involved in any significant two-way interaction, except that with Type, the factor which, by contrast, was involved in most of the significant interactions, both two-way and three-way. All interactions above three-way interactions were not significant. Note that a separate series of univariate tests showed that all significant main effects and interactions reported here were based on both BAR and FAA, although, in terms of effect sizes, the FAA measure generally contributed less to the effects than the BAR measure.

Table 3: Summary of relevant interaction effects of the multivariate general linear model based on the four fixed factors Medium, Type, Target, and Sample and the additional covariate Group.

	F	*df*1	*df*2	p	η_p^2
Medium*Group	0.916	2	46	n.s.	0.038
Medium*Sample	0.159	2	46	n.s.	0.007
Medium*Target	2.803	2	46	n.s.	0.109
Type*Group (Fig. 9)	5.713	14	34	< 0.001	0.702
Type*Sample (Fig. 8)	54.386	14	34	< 0.001	0.957
Type*Medium (Fig. 5)	8.116	14	34	< 0.001	0.769
Type*Target (Fig. 6)	232.057	14	34	< 0.001	0.989
Target*Sample (Fig. 7)	22.582	2	46	< 0.001	0.495
Type*Medium*Sample	6.906	14	34	< 0.001	0.739
Type*Medium*Target	8.503	14	34	< 0.001	0.778
Type*Target*Sample	40.587	14	34	< 0.001	0.944
Type*Target*Group	5.843	14	34	< 0.001	0.706

Separate multiple-comparisons tests showed that the BAR and FAA measurements made in the REF conditions differed significantly neither between the experimental sessions (linked to the within-subject factors) nor between the German and Danish participants (linked to the between-subjects factor Sample). That all participant samples and sessions showed statistically the same BAR and FAA brain activity in the REF context means that all significant differences outside the REF condition represent valid effects of the hate-speech stimuli rather than artifacts of different baseline levels of brain activity. On this basis, we present key aspects of the results received for the individual factors in more detail in the following subsections. All results reported in these subsections came out significant in the multiple-comparisons tests.

3.1 Medium

Figure 5 shows the Type*Medium interaction or, in other words, how the difference between written and spoken hate speech affected the BAR and FAA measurements across the types of hate-speech stimuli. The most obvious result is that the spoken presentation mode resulted in more extreme measurements, i.e. in more pronounced reactions of the participants to the stimuli than the written presentation mode. For example, the FGL, IMP, and HOL stimuli all caused a significantly higher arousal and a significantly stronger negative valence (all $p < 0.001$) when being heard than when being read. In the opposite direction, the RQ and especially the IRO stimuli were perceived as significantly less arousing and negative (all $p < 0.001$) when being heard than when being read. Note that the FAA values of the spoken IRO stimuli on average exceed 1.0, which indicates that the stimuli did not have a particularly negative effect on the participants. In fact, the FAA values created by the IRO condition were so high that they did not differ significantly from those of the zero-stimulus reference condition REF. The FAA values of the IND stimuli on average also exceeded 1.0 in the spoken presentation mode but were still significantly below the FAA values of both IRO and REF.

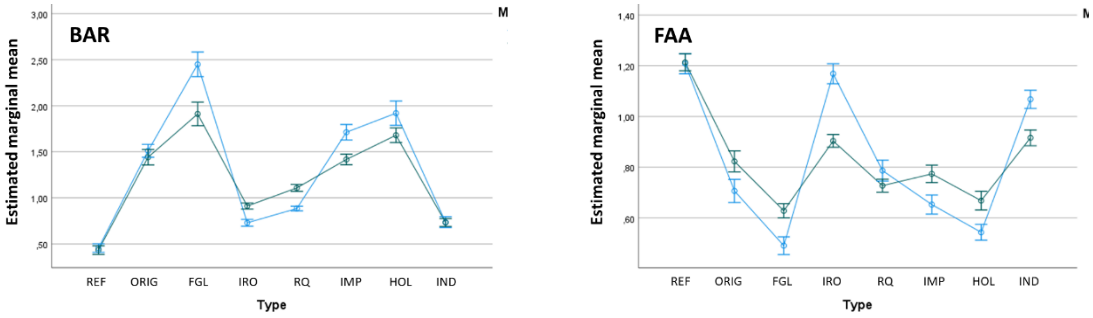

Figure 5: Results summary of the significant Type*Medium interaction (cf. Tab.3). Displayed are the estimated marginal means and their associated error bars (representing the 95 % CIs) of each stimulus type (across all Target and Sample conditions) for the spoken (blue) and the written stimuli (green). BAR (Arousal) results are shown in the left and FAA (Valence) results in the right panel. N = 100 per data point.

For some hate-speech types, Medium only significantly affected one of the two EEG measures. For example, in the case of ORIG and IND, only the FAA values differed significantly, with the spoken stimuli creating lower (ORIG) or higher (IND) FAA levels, respectively.

3.2 Type

Regarding Type, Figure 6 shows first of all that the REF condition yielded significantly lower BAR and higher FAA values than all other conditions (all $p < 0.001$). Furthermore, we see clear differences between the seven types of hate-speech. However, not all types of hate speech differ significantly from each other. For BAR, for example, the multiple-comparisons tests showed that the overall strongest arousal was caused by the FGL stimuli, followed by the HOL stimuli, which in turn were followed by the dyad of ORIG and IMP stimuli whose BAR measurements did not differ significantly from each other. The same applies to the dyad of IRO and RQ stimuli whose measurements were again lower than for the ORIG-IMP dyad. The significantly lowest BAR level and hence the lowest arousal was triggered by the IND stimuli.

For the FAA values, the seven types of hate speech form slightly different clusters in the multiple-comparisons tests. FGL and HOL both had the strongest negative effect on participants (i.e. they yielded the lowest FAA values), followed by the ORIG and IMP stimuli that formed a triplet together with the RQ stimuli. The dyad of IRO and IND stimuli together had the least negative impact or – given that FAA values are close to or higher than 1.0 – even had an ambivalent or slightly positive effect on participants.

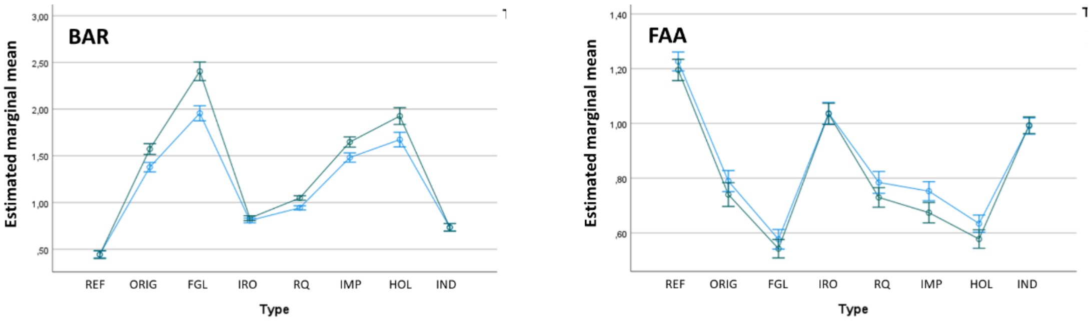

Figure 6: Results summary of the significant Type*Target interaction (cf. Tab.3). Displayed are the estimated marginal means and their associated error bars (representing the 95 % CIs) of each stimulus type (across all Medium and Sample conditions) for the foreigner- (blue) and Muslim-directed stimuli (green). BAR (Arousal) results are shown in the left and FAA (Valence) results in the right panel. N = 100 per data point.

With respect to the interaction of Type with Target, Figure 6 shows additionally that some types of hate speech triggered clear differences between stimuli addressing foreigners and stimuli addressing Muslims, whereas other types of hate speech did not. Overall, the pattern that emerged from the results was as follows: The stronger the impact of a stimulus type on participants (in terms of a higher arousal and a more negative valence), the more pronounced was the effect of Target. This is also supported by correlations (Pearson's r) that we calculated across the written and spoken stimuli (N=14) per stimulus type. The level of BAR was significantly positively correlated with the size of the difference between foreigner-directed and Muslim-directed hate speech (r[12] = 0.85, $p < 0.001$). For the FAA values, the correlation was weaker, yet also positive and significant (r[12] = 0.53, p =0.05). Accordingly, FGL and HOL stimuli triggered strong and significant differences between foreigner- and Muslim-directed hate speech, whereas IRO and IND stimuli did not.

3.3 Target

Beyond what was described in connection with the covariation between Type and Target in 3.2, we see in Figure 7 that Muslim-directed hate speech was significantly worse for participants. Compared to foreigners-direct hate speech, Muslim-directed hate speech caused a significantly higher level of arousal (BAR) as well as a significantly stronger negative valence (FAA). In addition, Figure 7 shows in terms of the interaction of Target with the factor Sample that Danish participants showed significantly stronger differences between the two target groups of foreigners and Muslims. In the case of BAR, this manifests itself in a higher arousal for Muslim-targeted hate speech compared to the German participants ($p < 0.001$). In the case of the FAA, the Danes perceived the foreigner-directed hate speech as less negative than the German participants ($p < 0.001$).

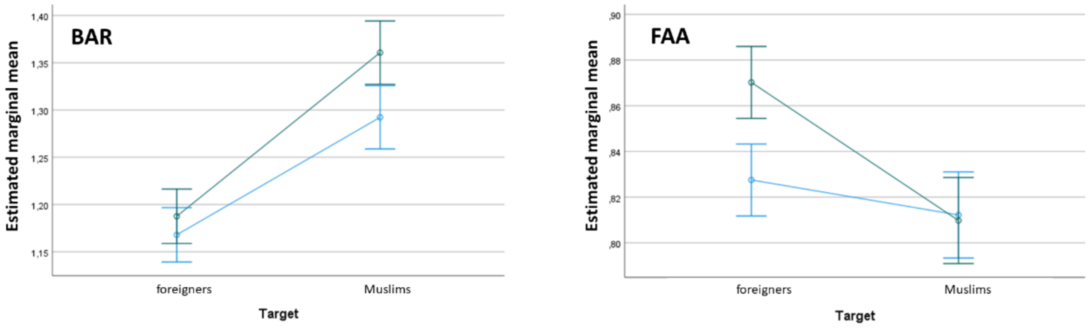

Figure 7: Results summary of the significant Target*Sample interaction (cf. Tab.3). Displayed are the estimated marginal means and their associated error bars (representing the 95 % CIs) of each targeted hate-speech group (across all Type and Medium conditions) for the German (blue) and Danish recipients (green). BAR (Arousal) results are shown in the left and FAA (Valence) results in the right panel. N = 400 per data point.

3.4 Sample

Figure 8 shows very clearly why the factor Sample produced, in terms of η_p^2, the weakest main effect but at the same time a particularly strong interaction effect with the factor Type. The main effect is based on the overall higher BAR values and FAA values obtained for the Danes compared to the Germans. In other words, the Danes were generally more aroused by hate speech than the Germans ($p < 0.01$), but the Germans perceived a generally more negative valence than the Danes ($p < 0.05$), cf. also Figure 7.

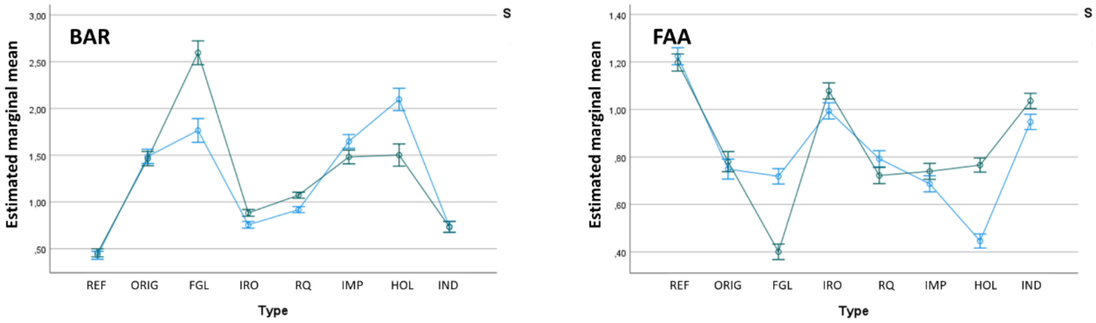

Figure 8: Results summary of the significant Type*Sample interaction (cf. Tab.3). Displayed are the estimated marginal means and their associated error bars (representing the 95 % Cis) of each stimulus type (across all Medium and Target conditions) for the German (blue) and Danish sample (green). BAR (Arousal) results are shown in the left and FAA (Valence) results in the right panel. N = 100 per data point.

Regarding the Sample*Type interaction, Figure 8 shows marked differences between the stimulus types. Interestingly, they encompass all stimulus types except for one: ORIG. However, also the IMP and IND differed only marginally (yet significantly, $p < 0.05$) between the German and Danish participants. The greatest differences related to Sample were obtained for the FGL and HOL stimuli. The Danes reacted most intensely to the FGL stimuli in terms of both arousal and negative valence ($p <0.001$). For the German participants, by contrast, the FGL stimuli were somewhere in the middle between ORIG and IMP with regard to arousal and valence. The results for the HOL stimuli turned out exactly the other way round. The Germans showed by far the most intense reaction to this type of stimulus, i.e. the highest arousal and the most negative valence ($p <0.001$), whereas for the Danes, the HOL stimuli were merely on a par with the ORIG and IMP stimuli – and in terms of FAA even on a par with the RQ stimuli.

Also note that for the IRO stimuli, the Danes reacted with a higher arousal and a less negative valence than the Germans ($p <0.01$).

3.5 Group

The analysis of the covariate Group led to a remarkable result. The five groups of people covered by the participant sample fell into two classes. This bipartition of the participant sample was identical for the Danish and the German sample and affected both measures BAR and FAA, which is why there was no significant interaction between Group and Sample.

One of these classes included three groups of participants who generally reacted more intensely to the hate speech stimuli, i.e. showed a relatively high arousal and a relatively strong negative valence. These three groups of participants were the pensioners, students and church employees. The other class was constituted by the two remaining groups of participants, lawyers and teachers. As illustrated in Figure 9, the lawyers and teachers reacted significantly less sensitively to hate speech stimuli ($p < 0.01$), the latter group even less than the former. This concerned in particular the ORIG, FGL, IMP, and HOL stimuli, i.e. precisely those stimulus types that had the strongest overall impact on the participants.

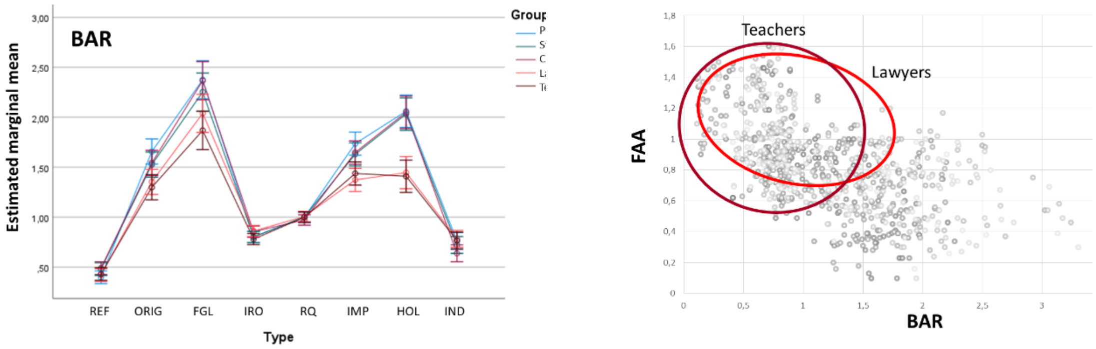

Figure 9: Left panel: results summary of the significant Type*Target interaction (cf. Tab.3) in terms of estimated marginal means and their associated error bars (representing the 95 % CIs) of each stimulus type for the five different groups in the German and Danish samples. N = 100 per data point. Right panel: range and distribution of BAR (Arousal) and FAA (Valence) measurements across all stimuli and participants of the German sample (N = 800).

Beyond the effect of the covariate Group, Figure 9 illustrates by means of the German sample that the measurements made for both BAR and FAA in the experiment were overall relatively evenly distributed over a large range of values – and, in addition, significantly negatively correlated with each other (in terms of Pearson's r). This correlation was somewhat stronger for the German sample shown in Figure 8 ($r [348] = 0.62$, $p <0.001$) than for the not shown Danish sample ($r [348] = 0.39$, $p <0.001$). That is, across all factors and conditions, the FAA values decreased the more the BAR values increased; or, to put it another way, the more a stimulus managed to arouse the participants, the more negative was also the valence that it was able to trigger. For hate speech stimuli, this is a plausible correlation.

4 Discussion

The present experiment investigated by means of two EEG measures, BAR and FAA (averaged per participant over the approx. 30 seconds of a stimulus condition), how samples of German and Danish participants perceived and evaluated hate-speech stimuli. In addition to authentic hate-speech stimuli (ORIG) from the XPEROHS Corpus, six other types of hate speech were tested. These were derived from the ORIG stimuli and, according to empirical evidence, represent the six most common morphosyntactically and/or prosodically marked types of hate speech in German and Danish (Neitsch and Niebuhr, 2021): figurative language (FGL), irony

(IRO), rhetorical questions (RQ), imperatives (IMP), Holocaust references (HOL) and indirectness (IND). All stimuli dealt with xenophobic hate speech. In that, half of the stimuli aiming at foreigners in general and the other half at Muslims in particular. In addition, all stimuli were presented to the participants in both written and spoken form, but in separate sessions. The samples of the German and Danish participants included 25 people each, consisting of five equally large subgroups: pensioners, lawyers, church employees, students and teachers. On the basis of this experimental framework, five research questions were addressed. In the following, we discuss our results in the light of these questions.

4.1 Question (I): Heterogeneity

Do the biosignals obtained mirror our earlier explicit-rating results by also forming a reaction continuum for the evaluation of the hate speech stimuli, for example, in terms of differences in the levels or amplitudes of EEG measurements? Based on our results, this question can be answered with a clear 'yes'. All results Figures 5-9, but especially the scatter plot in Figure 9, show that hate speech is not a homogeneous phenomenon in the domain of biosignals either. Not every type of hate speech was perceived as equally severe. Neither did every reader or listener perceive a certain stimulus as equally severe, see also the explanations on questions (II) and (V) below. Furthermore, severity can be operationalized on the basis of various measures and perceptual qualities. In the present study, these were arousal and valence, represented by the BAR and FAA parameters of the EEG signal. In our earlier studies, we operationalized severity via the two continuous rating scales "degree of personal unacceptability" and "strength of consequences for the originator".

For many cognitive processes and perceptual qualities such as pain (Gentile et al., 2011), creativity (Carroll et al., 2009), personality (Roberts and Woodmann, 2017), and charisma (D'Errico et al., 2013) there are empirically-based, tried-and-tested evaluation concepts. So far, there is nothing comparable for hate speech – most likely because both the (societal/juridical) discussion of hate speech as well as its treatment by social-media companies was dominated by the question of what hate speech is and what it is not. Such a binary perspective on hate speech has, in our opinion, implicitly promoted the view that all cases of hate speech are equivalent. This is a failure. Empirical evidence like that in the present study clearly shows that the perception of hate speech is not homogeneous, but varies depending on factors internal and external to the stimulus. A separate line of research is urgently needed to develop a concept for a multidimensional, standardized, gradual measurement of the perception of hate speech. Implicit and explicit parameters could interlock to that end, for example, by starting to determine relevant (sensitive) biosignal parameters. Results obtained with these parameters would then be associated with cognitive processes and perceptual qualities and, from there, translated back into attributes and scales for explicit ratings that are easier and more extensive to use in everyday use than biosignals. That is, the ultimate goal should be a test concept of explicit scales, but developed from a solid foundation of biosignal data.

One of our reviewers suggested that the impoliteness framework of Culpeper and colleagues could serve as a point of departure for such a test concept, or may even serve as an off-the-shelf system for guiding the analysis and classification of our data. The impoliteness framework is successfully applied for many years in the research fields of pragmatics and discourse analysis, see the overview paper by Culpeper et al. (2016). We have summarized the framework schematically in Figure 10. Like the reviewer, we also see overlaps of this framework with the types of German and Danish hate speech addressed here. Toning down a statement via an ironic prosody, for example, is also a strategy in the impoliteness framework of Culpeper and colleagues (referred to as "Sarcasm and Mocking"). Similarly, the use of slurs, i.e.

figurative language in our terms, falls under the “bold-on-record impoliteness” type in the framework of Culpeper and colleagues; and yet, whether the framework as a whole can be applied to hate speech, or whether all phenomena subsumed under hate speech are properly covered by the framework, these are questions that we would rather put up for discussion in future studies. There are several reasons for this.

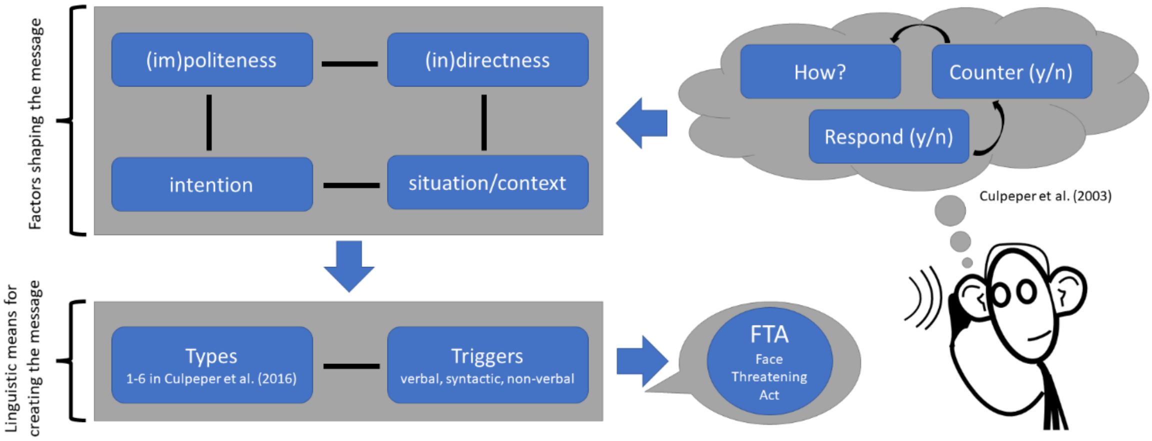

Figure 10: Schematic representation of the impoliteness framework of Culpeper and colleagues, see, for example, the overview in Culpeper et al. (2016).

First, the impoliteness framework is based on “face-threatening acts”, with "face" referring to a person's self-esteem or self-image. Person A performs a face-threatening act towards person B, see Figure 10. Given the empirical data on which the framework relies (see Culpeper et al. 2003), this happens primarily via the medium of spoken language and such that the face-threatening act takes place “in the hearing of the target” (Culpeper et al. 2016:437). B therefore in principle has the option of reacting or not reacting to this act, as is shown in Figure 10. However, such a direct discourse connection does not have to exist in the case of hate speech. On the contrary, we detail in Neitsch et al. (2021) that one of the main reasons for the rapidly growing problem of hate speech is the temporal and medial distance between the hate speaker and the target, at least in the typical written social media context of hate speech.

Second, continuing with this argument, hate speech does not necessarily mean that A speaks (impolitely) with B. It can also mean that A speaks (impolitely) about C while being in an interaction with B. In the case of impoliteness, the recipient is at the same time also the target (Fig. 10); this does not always apply to hate speech. The consequence of recipient and target not having to be identical is that hate speech can even be used to enhance the faces of A and B (at the expense of C), if A and B share the values expressed by their hate speech. This raises doubts about whether hate speech must always be a face-threatening act – and whether face is a useful concept at all if there is nobody directly involved who can lose his/her face. All of the stimuli evaluated by our participants are of this nature (A talks with B about C), but, of course, there are cases where hate speech is directly aimed at a reader or listener involved in the discourse.

Third, by definition, hate speech includes statements that seek to discriminate people with regard to their religion, ethnicity, nationality, race, color, descent, gender or other identity factors (see footnote 1). Thus, it is more an identity-threatening than a face-threatening act; and this makes hate speech, unlike impoliteness, a linguistic tool by which in-groups define themselves via out-groups. There is an inherently political or social component to hate speech; and because it typically targets minorities (as in our stimuli), hate speech occurs perhaps more often than not in an imbalanced power context in which the or originators of hate speech subjectively see themselves in a higher power position than their targets. By contrast, the examples of impoliteness presented by Culpeper and colleagues (cf. Culpeper et al. 2003) primarily take place in an inverse power relationship, i.e. they are addressed towards a higher-power (e.g., governmental) authority.

So, overall, there are good reasons in our opinion to differentiate between hate speech and impoliteness phenomenologically and conceptually – and thus also analytically and typologically. Nevertheless, we also see it as a worthwhile task for future studies to explore the fit of hate speech into the impoliteness frameworks of pragmatics research, and to draw inspiration from these frameworks for the analysis and classification of hate speech, especially with regard to the many verbal, syntactic, and nonverbal impoliteness triggers identified by pragmatics studies, some of which are discussed in Culpeper et al. (2016). Testing and quantifying their perceived severity in different contexts will yield results that can also fruitfully inform impoliteness frameworks and other developments in the field pragmatics.

4.2 Question (II): Cross-language differences

Do we find the same German-Danish differences in the evaluation of hate speech with regard to the target group (foreigners vs. Muslims) and the type (e.g., HOL vs. FGL)? Yes, in fact the arousal and valence results from the EEG signal reflect key aspects of previous explicit target and type ratings remarkably well. In these previous studies, the Danes reacted most sensitively to the FGL stimuli and the Germans to the HOL stimuli, see Figure 2. In addition, the Danes differentiated more than the Germans between foreigner and Muslim-oriented hate speech stimuli, for example, by rating especially the latter stimuli significantly worse than the Germans. Both result patterns manifested themselves also in our BAR and FAA measurements of arousal and valence.

Note in this context that the stronger differentiation between foreigner- and Muslim-directed hate speech among the Danes' explicit ratings primarily concerned the y-axis of the 2D rating space (see Fig.2). In the EEG signal, it was mainly arousal (BAR) that showed the stronger reaction of the Danes to the Muslim-directed hate speech. Furthermore, the explicit ratings of the Germans were overall further to the right on the x-axis than those of the Danes, indicating an overall higher degree of personal unacceptability (see Fig.2). In the EEG signal, we found an overall significantly lower FAA level for the Germans, i.e. a significantly more negative valence as compared to the Danes. These parallels suggest a relationship between the explicit and implicit (EEG) ratings. More specifically, the "degree of personal unacceptability" seems to be reflected more in the valence results (FAA measurements) and the "strength of consequences for the originator" more in the arousal results (BAR measurements). It is important to investigate such possible correspondences between explicit and implicit ratings further in future studies – not least with a view to the idea of an evaluation concept for hate speech rooted in biosignal research, see 4.1. Additional biosignals (like breathing patterns, skin-conductance response, and pupillometry) have to be taken into account by these future investigations, as well as additional rating-scale labels. What we need is a better understanding of whether and how explicit rating scales (or their labels) can be mapped onto biosignal parameters.

In connection with such a mapping, it will further be informative to use the reaction times of explicit ratings or, in particular, the temporal shapes of biosignals to determine more precisely which linguistic elements in the stimuli trigger the participants' evaluations to what degree. Because of their very short and clearly identifiable stimulus-to-response latencies, EEG signals are suitable to that end, but pupil-dilation signals even more so, although they mainly reflect the arousal and not the valence dimension. See Johansson and Balkenius (2018) for how emotionally charged stimuli affect pupil dilations and response latencies. An underlying assumption in our study, as well as in all previous studies with explicit hate-speech ratings (cf. 1.2), is that participants respond to the stimuli as a whole. One of our reviewers pointed out that, in principle, our data does not allow us to differentiate between participants who reacted negatively to the hate speech messages as such and participants who were only bothered by the target-group keywords Muslims and foreigners. The significant effects of stimulus type and medium alone speak against this possibility. In addition, in a recently published study with a baseline condition (Niebuhr, 2022), it was shown that target-group terms per se (embedded in otherwise neutral statements) do not trigger an evaluation as strong as that of real hate speech. Nevertheless, the reviewer's comment highlights the importance to exploit, in a following step, the temporal resolution of signals for a more detailed understanding of how individual linguistic elements contribute to the overall evaluation of hate speech stimuli – and to work out inter-individual differences in that context.

4.3 Question (III): Written vs. spoken hate speech

Do we find supporting evidence for the relevance of the hate-speech medium, for example, in terms of some hate-speech types causing stronger physiological reactions in the spoken domain (HOL, IMP, FGL) and other types in the written domain (IRO, RQ)? Once again, the EEG data of the present study allow us to answer this question with 'yes'. The biosignals of arousal and valence were consistent with previous findings according to which morphosyntactically marked hate speech such as FGL, IMP, HOL becomes worse when it is spoken our loud and not just read silently; and that the exact opposite applies to IRO and RQ hate speech whose communicative function is primarily marked prosodically. Our results additionally showed with respect to question (III) that the valence (FAA) measure was overall a little more sensitive in capturing the effects of the presentation medium than the arousal (BAR) measure (cf., e.g., ORIG and IND). In the explicit ratings of our previous studies, the effects of medium were also more strongly associated with one of the two measures: that of the x-axis, see Figure 2. This parallel corroborates our above assumption (see 4.2) that valence-related hate speech evaluation is more closely linked to the "degree of personal unacceptability" ratings (and arousal-related hate speech evaluation more to the "strength of consequences for the originator" ratings).

4.4 Question (IV): The role of prosody

Do we find indications that prosody is able to attenuate hate speech to such an extent that it no longer has to be classified as hate speech? The present study was not designed to define an objective, parametric threshold value for the classification of stimuli into hate speech and non-hate speech. Nonetheless, our study included the zero-stimulus reference condition REF, in which participants simply sat in their comfortable chair in a quiet, familiar environment. Additionally, the REF condition was recorded before the participants saw or heard any hate-speech stimulus. It is therefore reasonable to assume that participants experienced this condition as neutral or even slightly positive, as far as this was possible in an experimental setting. Average FAA values of consistently > 1.0 support this assumption. Given that, it is all the more remarkable that the spoken IRO stimuli on average triggered just as little a negative reaction as the REF condition – in both the German and especially the Danish participants. That is, it was statistically indistinguishable in EEG terms whether participants sat quietly in a comfortable chair or listened to ironic hate speech. This was true only for the valence dimension of the EEG signal (FAA), which is, however, perhaps more relevant for deciding what hate speech is and what is not than the arousal dimension (BAR).

Irony is a type of hate speech that is primarily marked prosodically. It is plausible in view of this fact that the valence response was less positive to the written than to the spoken stimuli.

Nevertheless, even the written IRO stimuli, in which part of the ironic prosody was spelled out orthographically, achieved valence reactions close to an FAA value of 1.0. Against this background, the answer to question (IV) is 'yes'. It seems possible that the use of ironic prosody is indeed be able to turn hate speech into non-hate speech – especially spoken hate speech. Recall in connection with this conclusion that the IRO stimuli were derived from the authentic ORIG stimuli, which were evaluated as being significantly worse than the IRO stimuli in all (implicit and explicit) respects. Given this considerably discrepancy to the ORIG stimuli, irony seems to be a very effective tool to attenuate hate speech content.

This is a potentially momentous conclusion, for one thing, because it would make the task of automatically identifying and deleting hate speech considerably more difficult, particularly based on key words and key phrases alone, as irony essentially relies on context; and for another thing, because it would give prosody a whole new role in the definition and classification of hate speech, i.e. a layer of linguistic structure and expression that is largely absent in the graphemic representation of written language and hence, for this primary hate-speech medium, only exists in the minds of authors and recipients. In view of this, it is immensely important that future research systematically investigates the concept of “implicit prosody” (Fodor, 2002) for the domain of hate speech, see Niebuhr (2022).

Note that our conclusion about the attenuating effect of irony in hate-speech evaluation is limited by an important aspect. The participants in our experiment did not belong to the target groups addressed in the stimuli. We did not test any foreigners or Muslims. Whether irony still has such a strong attenuating effect when the recipients belong to the target group addressed in the stimuli will be a question with high priority for follow-up studies. Initial data suggests that it does matter for the evaluation of the stimuli whether participants do or do not have a migration background and belong to the actual addressees of hate-speech messages (Niebuhr 2022).

4.5 Question (V): The role of the individual

Beyond the replication of previous findings, what additional insights do the EEG biosignals provide – for example, with regard to “socially desired responses” and inter-individual variation? The most notable result regarding inter-individual variation was undoubtedly the effect of the covariate Group. We actually compiled the German and Danish samples from different groups of participants in order to increase the representativeness of the experiment’s results. It was not expected that this compilation procedure would reveal group-specific differences in the evaluation of hate speech, even across the German and Danish samples. Our study does not provide any clear indications as to why lawyers and teachers reacted significantly less sensitively to hate speech than church employees, students, and pensioners. Table 2 shows no obvious differences between lawyers and teachers on the one hand and church employees, students, and pensioners on the other. For example, if age were the critical factor, for example, in the form of a dividing line between younger participants reacting differently to hate-speech stimuli than older participants, then teachers and students (i.e. the two youngest groups of people) would have behaved similarly, but not teachers and lawyers. We suspect, however, that our questions about familiarity and experience with hate speech were too imprecise to reveal how exactly the lawyers and teachers differed from the other groups. It is reasonable to assume that, firstly, lawyers and teachers deal with hate speech much more frequently than the other groups. Secondly, lawyers and teachers deal with hate speech not only and perhaps not even primarily in their free time. Rather, dealing with hate speech is (also) part of their everyday professional life. It is therefore to be expected that lawyers and teachers can (and must) deal with hate speech stimuli in a more objective, distant, and factual manner than the other three groups of people. Of course, this just an assumption and not a conclusion. What corroborates this assumption is that the teachers and lawyers stood out in a similar ways from the other three societal groups in two entirely independent participant samples, i.e. those of the two languages German and Danish. In numbers, this means that 2 x 2 x 5 people behaved differently in their EEG signals than the other 2 x 3 x 5 people. This in no way rules out mere coincidence, but it does make it less likely.

It would be worthwhile to examine in subsequent studies whether this assumption can be further supported and, if so, whether general hate-speech coping strategies can be developed on this basis for the society and/or vulnerable individuals.

With regard to “socially desired responses”, there are major and minor deviations to be discussed between the explicit ratings and the measured BAR and FAA values. One of the minor deviations concerns the HOL stimuli. In the explicit ratings, the Danes considered them less severe than the FGL stimuli, but still more severe than other types of stimuli, such as ORIG, IMP, and RQ, especially on the y-axis that concerned the demanded "strength of consequences for the originator", see Figure 2. In the biosignals, however, the HOL stimuli did not stand out separately anymore in the Danish sample. Rather the HOL stimuli clustered together with the ORIG, IMP, and RQ stimuli. This indicates that Danes actually judge HOL stimuli to be less severe than their explicit ratings suggest.

Furthermore, the Danes were, according to their EEG biosignals, also (still) significantly more tolerant of ironic hate speech than the Germans. One could ask (and examine more closely in future studies) whether this greater tolerance has something to do with either the black humor for which the Danes are known worldwide (Levisen, 2018) – or with the high one status of freedom of speech in Denmark. As Nielsen (2019) states: "freedom of speech is highly regarded and protected in the constitution, which means that a foreign libel victim is unlikely to institute court proceedings in Denmark, even if the wrongdoer is domiciled in Denmark, because freedom of speech in most cases will take priority over defamation and privacy rights" (pp. 33-34).

The most important point regarding “socially desired responses” concerns the IND stimuli, though. The present study replicates and substantiates the corresponding findings from earlier studies on a larger empirical basis: In terms of hate-speech evaluation or severity, the biosignal reactions to IND stimuli were significantly weaker than the explicit ratings suggested, cp. Figures 2 and 8. That is, preceding an ORIG hate-speech statement with the phrase "I have nothing against __, but ___" can significantly reduce the statement’s hate-speech effect – at least in terms of what the recipients really think about the stimulus or its originator. The effect is not as strong as with irony; and yet phrases like “I have nothing against __, but ___” also seem to be an effective instrument to shift hate speech towards non-hate speech. Moreover, this is true independently of the addressed target group and applies even more so for Danish than for German recipients. Note, of course, the same critical limitation that we already stressed in connection with the IRO stimuli: Our samples did not involve any participants from the target groups addressed in the stimuli. We do not know if and how differently the real target groups react to the presented hate speech stimuli – both explicitly and in the terms of biosignals. Besides the IND and IRO stimuli, this applies also to the FGL and HOL stimuli, all of which produced particularly striking results in the present study. Thus, these four stimulus types should be prioritized in follow-up studies that involve the stimuli’s actual target groups.

With regard to measuring hate speech itself, the present biosignal data were able to refine previous result patterns obtained with the explicit ratings in some aspects. This includes the present finding that the effects of Target in terms the experienced differences between foreigner- and Muslim-oriented hate speech increased with increasing arousal and valence

(BAR and FAA) values. That is, some stimulus types (morphosyntactic or contextual frameworks) always triggered similarly weakly negative responses, regardless of the addressed target group. Other frameworks, by contrast, are already “charged” with hate speech to such an extent that their perceived severity increases significantly if they are additionally combined with a politically or socially sensitive and/or very specific target group like Muslims as compared to foreigners.

5 Summary and outlook

In summary, we return to the comparison made in the introduction between hate speech and bistable images. The present findings are a loud and strong plea that the recipient must not be ignored in connection with hate speech. Trying to operationalize hate speech via the stimulus or the signal alone will never be entirely successful or reliable. This applies to hate speech identification, but even more so to its evaluation. Germans react differently than Danes to morphosyntactically comparable hate speech stimuli, and even some people or professional groups react differently than others to identical hate speech stimuli. Different types of hate speech can also cause very different reactions. IND and IRO stimuli, for example, can be based on the same key words or key phrases and aim at the same target groups as HOL and FGL stimuli; and yet, recipients would react entirely different to them.

We have to research and understand hate speech from the perspective of the recipient in order to arrive at a reliable and sensitive identification and evaluation. This is all the more true when it comes to the question of social or legal consequences for the originator. Fair consequences, in particular, can only be achieved through a better understanding of the recipient's perspective. In addition, as our results show, such a perspective must be based on biosignals and not on explicit ratings in order to avoid bias or artifacts due to “socially desired responses”. This also means that current approaches to the automatic identification of hate speech are to some extent inadequate, because they are based on explicit ratings of, rather than on biosignals responses to hate speech stimuli.

The priorities for the following steps of our line of research are on the following three points: (1) a better understanding of how biosignals relate to explicit ratings or scale labels, so that simple and more precise test procedures for hate speech can be developed; (2) the replication of the study with the actual target groups of the hate-speech stimuli, i.e. foreigners in general and Muslims in particular, focused on the FGL, HOL, IRO and IND stimuli, but with taking both explicit ratings and biosignals measures; (3) the attempt to define a measurable threshold for the classification of stimuli as hate speech and non-hate speech, based on multidimensional and multimodal biosignals.

6 Acknowledgments

The XPEROHS project (project number 95-16416) was kindly funded by the Velux Foundation. Furthermore, we are greatly indebted to Albert Rilliard, our other anonymous reviewer and our editor, Plinio Barbosa, for their questions and comments that helped a lot to improve the manuscript and will be useful for us in future studies as well. Further thanks are due to Bente Olsen, Klaus Geyer, Eckhard Bick, Martin Carlsholt Unger, Andrea Kleene and Rasmus Nielsen for many inspiring and fruitful discussions about the nature of hate speech and, in the case of Martin, also for his help in recruiting Danish participants. Finally, we would like to thank Anna Gutnyk for allowing us to publish the photograph that we took of her while wearing the MUSE II EEG headset.

REFERENCES

1. Asif A, Majid M, Anwar SM. Human stress classification using EEG signals in response to music tracks. Computers in Biology and Medicine. 2019; 107: 182-196.

2. Ateş C, Kaymaz Ö, Kale HE, Tekindal MA. Comparison of Test Statistics of Nonnormal and Unbalanced Samples for Multivariate Analysis of Variance in terms of Type-I Error Rates. Comput Math Methods Med. 2019:8.

3. Balcerzak B, Jaworski W. Application of linguistic cues in the analysis of language of hate groups. Computer Science. 2015; 16.

4. Baumgarten N, Bick E, Geyer K, Iversen DA, Kleene A, Lindø AV, ... Petersen EN. Towards balance and boundaries in public discourse: expressing and perceiving online hate speech (XPEROHS). International Journal of Language and Communication. 2019; 50: 87-108.

5. Beer J, Beer J, Markley RP, Camp CJ. Age and living conditions as related to perceptions of ambiguous figures. Psychol. Rep. 1989; 64: 1027–1033.

6. Bialystok E, Shapero D. Ambiguous benefits: The effect of bilingualism on reversing ambiguous figures. Developmental Science. 2005; 8: 595-604.

7. Bick E. An Annotated Social Media Corpus for German. Proc. 12th International Conference on Language Resources and Evaluation, Marseille, France. 2020: 6127-6135.

8. Bick E, Geyer K, Kleene A. „Die ách so friedlichen Muslime “: Eine korpusbasierte Untersuchung von Formulierungsmustern fremdenfeindlicher Aussagen in Sozialen Medien. In Wachs S, Koch-Priewe B, Zick, A, editors. Hate Speech-Multidisziplinäre Analysen und Handlungsoptionen. Wiesbaden: Springer; 2021. pp. 81-103.

9. Bregman, AS. Auditory scene analysis: The perceptual organization of sound. Cambridge: MIT press; 1994.

10. Calderón FH., Balani N, Taylor J, Peignon M, Huang YH, Chen YS. Linguistic Patterns for Code Word Resilient Hate Speech Identification. Sensors. 2021; 21: 7859.

11. Cao T, Wang L, Sun Z, Engel SA, and He S.. The independent and shared mechanisms of intrinsic brain dynamics: Insights from bistable perception. Frontiers in Psychology. 2018; 9: 589.

12. Carroll EA., Latulipe C, Fung R, Terry M. Creativity factor evaluation: towards a standardized survey metric for creativity support. Proceedings of the seventh ACM conference on Creativity and cognition. 2009: 127-136.

13. Culpeper J. Impoliteness Strategies. In Capone A, Mey J L, editors. Interdisciplinary Studies in Pragmatics, Culture and Society (Vol. 4, pp. 421–445). Berlin: Springer International Publishing; 2016.

14. Culpeper J, Bousfield D, Wichmann A. Impoliteness revisited: With special reference todynamic and prosodic aspects. Journal of Pragmatics. 2003; 35: 1545–1579.

15. Dattalo P. Analysis of Multiple Dependent Variables. Oxford: Oxford University Press; 2013.

16. Davidson T, Warmsley D, Macy M, Weber I. Automated hate speech detection and the problem of offensive language. Proc. 11th International AAAI Conference on Web and Social Media ICWSM '17. 2017: 512-515.

17. D'Errico F, Signorello R, Demolin D, Poggi I. The Perception of Charisma from Voice: A Cross-Cultural Study. Proc. Humaine Association Conference on Affective Computing and Intelligent Interaction. 2013: 552-557.

18. Fodor, JD. Psycholinguistics Cannot Escape Prosody. Proc. 1st International Conference on Speech Prosody, Aix-en-Provence, France. 2002: 83–88.

19. Fortuna P, Nunes S. A survey on automatic detection of hate speech in text. ACM Computing Surveys (CSUR). 2018; 51: 1-30.

20. Gale AG, Findlay JM. Eye movement patterns in viewing ambiguous figures. Eye movements and psychological functions: International views. 1983: 145-168.

21. Gambäck B, Sikdar UK. Using convolutional neural networks to classify hate-speech. Proc. 1st Workshop on Abusive Language Online. 2017: 85-90.

22. García-Acosta A., Riva-Rodríguez JDL, Sánchez-Leal J, Reyes-Martínez RM. Neuroergonomic Stress Assessment with Two Different Methodologies in a Manual Repetitive Task-Product Assembly. Computational Intelligence and Neuroscience. 2021.

23. Garcia-Moreno FM, Bermudez-Edo M, Garrido JL, Rodríguez-Fórtiz MJ. Reducing response time in motor imagery using a headband and deep learning. Sensors. 2020; 20: 6730.

24. Gentile DA, Woodhouse J, Lynch P, Maier J, McJunkin T. Reliability and validity of the Global Pain Scale with chronic pain sufferers. Pain Physician. 2011; 14: 61-70.

25. Geyer K. Entmenschlichende Metaphern in ethnotroper („fremdenfeindlicher“) Hatespeech in sozialen Medien. In Bülow L, Marx K, Meyer-Vieracker S, Mroczynski, R, editors. Digitale Pragmatik. Heidelberg: J. B. Metzler; 2021.

26. Geyer K, Bick E, Kleene A. “I am not a racist, but …”. A Corpus-Based Analysis of Xenophobic Hate Speech Constructions in Danish and German Social Media Discourse. In Knoblock N, editor. Grammar of Hate: Morphosyntactic Features of Hateful, Aggressive, and Dehumanizing Discourse. Cambridge: Cambridge University Press; 2021.

27. Geyer K. Die ‚Grammatik‘ der Hassrede – am Beispiel des Dänischen. In Strässler J. editor. Sprache(n) für Europa. Mehrsprachigkeit als Chance. Frankfurt: Peter Lang; 2019. pp. 195-207.

28. Goldstein, EB. Blackwell handbook of sensation and perception. John Wiley and Sons; 2008.

29. Handel S. Listening: an Introduction to the Perception of Auditory Events. Cambridge: MIT Press; 1989.

30. Herman K, Ciechanowski L, Przegalińska A. Emotional well-being in urban wilderness: Assessing states of calmness and alertness in informal green spaces (IGSs) with muse—Portable EEG headband. Sustainability. 2021; 13: 2212.

31. Hrdina M. Identity, activism and hatred: Hate speech against migrants on Facebook in the Czech Republic in 2015. Naše společnost. 2016; 14: 38–47.

32. Jaki S, De Smedt T. Right-wing German hate speech on Twitter: Analysis and automatic detection. arXiv preprint arXiv:1910.07518. 2019.

33. Johansson B, Balkenius C. A computational model of pupil dilation. Connection Science. 2018;30: 5-19.

34. Klem GH, Lüders HO, Jasper HH, Elger C. The ten-twenty electrode system of the International Federation. The International Federation of Clinical Neurophysiology. Electroencephalography and Clinical Neurophysiology, Supplement. 1999; 52: 3–6.

35. Klintman H. Original thinking and ambiguous figure reversal rates. Bulletin of the Psychonomic Society. 1984; 22: 129-131.

36. LaRocco J, Le Minh D, Paeng D-G. A Systemic Review of Available Low-Cost EEG Headsets Used for Drowsiness Detection. Frontiers in Neuroinformatics. 2020; 14: 1-42.

37. Laukkonen RE, Tangen JM. Can observing a Necker cube make you more insightful? Consciousness and Cognition. 2017; 48: 198-211

38. Long GM, Toppino, TC. Multiple representations of the same reversible figure: Implications for cognitive decisional interpretations. Perception. 1981; 10: 231-234.

39. Levisen C. Dark, but Danish: Ethnopragmatic perspectives on black humor. Intercultural Pragmatics. 2018; 15: 515-531.

40. Malmasi S, Zampieri M. Detecting hate speech in social media. arXiv preprint arXiv:1712.06427. 2017.

41. MacAvaney S, Yao HR, Yang E, Russell K, Goharian N, Frieder O. Hate speech detection: Challenges and solutions. PloS one. 2019; 14: e0221152.

42. Martins R, Gomes M, Almeida JJ, Novais P, Henriques P. Hate speech classification in social media using emotional analysis. Proc. 7th Brazilian Conference on Intelligent Systems (BRACIS), IEEE. 2018: 61-66.

43. Neitsch J, Niebuhr O. Assessing hate-speech perception through bio-signal measurements: A pilot study. Proc. Biosignale 2020, Kiel, Germany, 2020: 66-67.

44. Neitsch J, Niebuhr O. On the role of prosody in the production and evaluation of German hate speech. Proc. 10th International Conference of Speech Prosody, Tokyo, Japan, 2020: 710-714.

45. Neitsch J, Niebuhr O, Kleene A. What if hate speech really was speech? Towards explaining hate speech in a cross-modal approach. In Wachs S, Koch-Priewe B, Zick, A, editors. Hate Speech-Multidisziplinäre Analysen und Handlungsoptionen. Wiesbaden: Springer; 2021. pp. 105-135.

46. Neitsch J, Niebuhr O. Types of hate speech: How speakers of Danish rate spoken vs. written hate speech. Proc. 4th International Conference of Phonetics and Phonology in Europe, Barcelona, Spain. 2021: 1-2.

47. Neitsch J, Niebuhr O. Die Erforschung geschriebener und gesprochener Hassrede im Deutschen: bisherige E-kenntnisse zu Prosodie und Kontext. In Jaki S, Steiger S, editors. Digitale Hate Speech - Interdisziplinäre Perspektiven auf Erkennung, Beschreibung und Regulation. Berlin: Springer; 2023.

48. Niebuhr O. Prosody in hate speech perception: A step towards understanding the role of implicit prosody. Proc. 11th International Conference of Speech Prosody, Lisbon, Portugal, 2022: 520-524.

49. Nielsen PA. Choice of Law for Defamation, Privacy Rights and Freedom of Speech. Oslo Law Review. 2019; 6: 32-42.

50. Papcunová J, Martončik M, Fedáková D, Kentoš M, Bozogáňová M, Srba I, ... Adamkovič M. Hate speech operationalization: a preliminary examination of hate speech indicators and their structure. Complex and Intelligent Systems. 2021: 1-16.

51. Peters MA. Limiting the capacity for hate: Hate speech, hate groups and the philosophy of hate. Educational Philosophy and Theory. 2020: 1-6.

52. Richer R, Zhao N, Amores J, Eskofier BM, Paradiso JA. Real-time mental state recognition using a wearable EEG. Proc. 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Honolulu, USA. 2018: 5495-5498.

53. Roberts R, Woodman T. Personality and performance: Moving beyond the Big 5. Current Opinion In Psychology. 2017; 16: 104-108.

54. Rodríguez-Martínez GA, Castillo-Parra H. Bistable perception: neural bases and usefulness in psychological research. International Journal of Psychological Research. 2018; 11: 63-76.

55. Ruwandika NDT, Weerasinghe AR. Identification of hate speech in social media. Proc. 18th International Conference on Advances in ICT for Emerging Regions (ICTer), IEEE. 2018: 273-278.

56. Shechter S, Hillman P, Hochstein S, Shapley RM. Gender differences in apparent motion perception. Perception. 1991; 20: 307–314.

57. Waseem Z, Hovy D. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. Proc. NAACL Student Res. Work. 2016: 88-93.

58. Zhang X, Bachmann P, Schilling TM, Naumann E, Schächinger H, Larra, MF. Emotional stress regulation: The role of relative frontal alpha asymmetry in shaping the stress response. Biological psychology. 2018; 138: 231-239.

59. Zhao G, Zhang Y, Ge Y. Frontal EEG asymmetry and middle line power difference in discrete emotions. Frontiers in behavioral neuroscience. 2018; 12: 225.