On the role of prosody in the production and evaluation of German hate speech

Jana Neitsch, Oliver Niebuhr

Centre for Industrial Electronics, MCI, University of Southern Denmark, Sønderborg, Denmark
neitsch@mci.sdu.dk, oniebuhr@mci.sdu.dk

Abstract

Hate speech targeting minority groups is a growing source of concern and not restricted to written language. It also occurs in spoken language in and beyond social media platforms. Given that, it is striking how little is known so far about the communicative and linguistic mechanisms of hate speech. The present study on German investigates participants' evaluation of subtypes of spoken hate speech (irony: IRO; Holocaust reference: HOL), derived from original (ORIG) hate speech items contained in a Twitter-Facebook corpus. The hate-speech items were elicited from a phonetically trained speaker and, in this spoken form, rated by listeners on two dimensions: personal (un)acceptability and necessity of legal/societal consequences for the speaker. Beyond correlations of these ratings with the prosody of the spoken hate speech items, we found lowest ratings for IRO and highest for HOL items, with ORIG items falling in between the two extremes. In conclusion, hate speech is no homogeneous phenomenon in terms of its perceptual evaluation and, in the case of spoken hate speech, prosody has an effect on how severely hate speech is rated.
Index Terms: hate speech, German, prosody, irony, holocaust, f0, tempo, voice quality.

1. Introduction

Hate speech is becoming more and more of a concern in societies around the world [1]. The term *hate speech* as such, however, remains highly controversial. There is a lack of consensus on its definition and impact, while the motivation and justification for its criminalization and regulation are inextricably linked to protecting freedom of speech on the one hand and the human rights of equality and dignity on the other [2]. Given the pressure that hate speech exerts on the pillars of modern civilization, it is striking how little is known about the linguistic and communicative mechanisms underlying the expression and perception of hate speech.

Addressing some of these gaps, the pilot perception study presented here is a part of the XPEROHS project. The project is based on a large sample of about 1.7 million real written hate speech tokens extracted from Twitter and Facebook posts in two languages: German and Danish (see [3,4] for further details, data, and goals of the project). Hate speech is primarily associated with written language, at least in the way it is experienced and discussed in the media [5]. However, hate speech, of course, also appears in spoken language, e.g., in political discourse [6] or in connection with bullying on school yards and football fields [7].

Our previous study showed that there is considerable prosodic variation among spoken hate-speech expressions [8] and that this variation is not primarily shaped by conveying hate speech itself. That is, we found no evidence that hate speech represents a communicative function in its own right with a separate prosodic feature setting like the recurrent, well established feature setting of, for example, contrastive focus [9]. Rather, it is the lexical and functional embedding of a hate-speech expression that shapes its prosodic characteristics. For instance, if a hate speech expression involves irony, then it is the signaling of irony that shapes the prosody of this hate-speech expression; and if a hate-speech expression involves a rhetorical question, then it is this specific prosodic construction [10,11] that determines how the hate-speech expression is realized prosodically.

The present study has three aims that concern the German data branch of the XPEROHS project. 1) We examine, for a subset of German hate-speech tokens, how diversely hate-speech expressions are rated (by German listeners) along two scales: *personal acceptability* and *culpable violation of societal norms*. The subset was selected such that its realization by a speaker would yield large prosodic differences between the items. On this basis, we test 2) to what extent prosodic-parameter variation can predict hate-speech ratings along the acceptability and norm-violation scales. 3) By comparing the degrees of this predictability between the selected lexical and functional embeddings of hate speech, we get an initial idea about the status of prosody in hate-speech ratings relative to other lexical and morpho-syntactic features.

2. Method

2.1. Design of the stimulus material

Twelve hate-speech items (i.e., posts) were selected from the German part of the Twitter and Facebook corpus. They were all similarly short. That is, they consisted of less than 25 words and included between 20 and 30 syllables. Moreover, semantically they were all directed against the minority group of immigrants.

From these 12 original hate-speech items (henceforth ORIG items), two further item sets were derived: One set that expressed irony (IRO items), and one set that included a Holocaust reference (HOL items). The IRO items were created, e.g., by prefixing the ORIG items with a phrase like *I would NEVER say that...* The HOL items were created by adding phrases such as *Throw them into a concentration camp!* to the end of the ORIG items. The IRO and HOL conditions were chosen because they emerged from the real Twitter and Facebook material as characteristic classes of German hate speech [3,12]. The way of creating the IRO and HOL items, i.e., by prefixing or appending phrases to the ORIG items, also took into account the natural differences between these three item classes in the real Twitter and Facebook data.

The three item classes were also chosen because their realization in spoken language would yield large prosodic differences. A preceding pilot production study [13] showed that, compared to ORIG items, HOL and IRO items are realized breathier (< HNR, dB) and at lower intensity (RMS, dB) and mean-f0 levels (Hz). IRO items are additionally realized at a slower speaking rate (syll/s) and with a larger f0 range (semitones) than HOL items. ORIG items show a higher value for the Hammarberg index (dB) of voice quality.

2.2. Elicitation of the stimulus material

The 12 items of the ORIG, IRO, and HOL conditions (i.e. the 36 items in total) were realized by a single male native speaker of (Northern Standard) German, BP. By being male, Caucasian, and between 35-50 years old, BP meets the typical profile of a hate speaker [7]. Moreover, BP is a professional speaker with a PhD degree in phonetics and linguistics. He is able to control the phonetic characteristics of his speech and to deliberately choose and produce phonetic patterns in order to create specific semantic-pragmatic effects. So, although only a single speaker was used to elicit the stimulus material of the present study, his professional background resulted in stimuli that are representative of emotional and expressive hate speech and clearly recognizable as such.

BP received the instruction to familiarize himself thoroughly with the 36 hate-speech items and to practice their elicitation with different phonetic realizations in order to find one realization that suits each individual item and makes it sound like authentic, spontaneously spoken hate speech. The speech-production task was conducted in the sound-proof booth of the Kiel Phonetics Lab [8]. Recordings were made with a Microtech Gefell M940 microphone at a 44.1 kHz sampling rate and a 16-bit quantization.

2.3. Design of the perception experiment

The experiment was based on two scales along which participants rated the stimuli. The scales referred to the scope of hate speech and the involved parties. The first scale asked participants to rate the degree to which the hate-speech item in question would be (un)acceptable for them personally. Thus, the participants indirectly reflected about if they could ever imagine using this hate-speech item themselves. The second scale required the participants to rate the extent to which the hate-speech item (if heard in an everyday situation) should have consequences for the speaker. This is a rating not driven by personal opinion but by higher inter-personal standards. Accordingly, consequences were defined in terms of legal actions or social reprisals due to a violation of general societal norms.

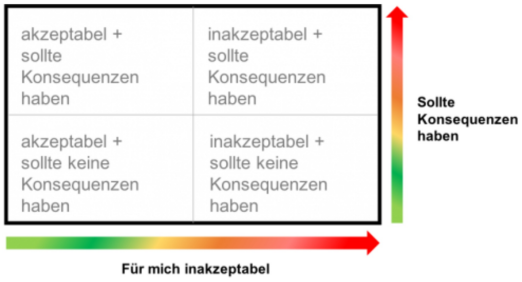

Figure 1: *2D rating space; x-axis: 'unacceptable for me', y-axis: 'should have consequences'. The grey texts were only present in the initial explanation of the 2D rating concept.*

As an innovative assessment method, a two-dimensional rating space was created, as is illustrated in Figure 1. The personal (non-)acceptability of hate-speech items was rated on the x-axis. The ratings on legal or societal consequences took place on the y-axis. So, with a single mouse click into the 2D rating space, both ratings were made simultaneously by participants. Compared to other rating paradigms, this innovative method yields simple, fast and intuitive responses. The dependent variables generated by this method are the respective x and y coordinates specifying where each item was placed by participants within the 2D rating space.

2.4. Experimental procedure

The perception experiment was run with a total of 28 participants (16 female, 10 male, 1 diverse, 1 NA, ∅ = 37.05 years, SD = 11.57 years). All participants were native speakers of German and naïve with respect to both the background of the research project and the linguistics of hate speech, including its various subtypes and conditions. In fact, all participants practiced normal non-academic professions.

The participants conducted the experiment individually on a laptop computer in a silent room that they knew well and felt comfortable in (e.g., their own living room or office). This was considered important in view of the partially disturbing content of the hate-speech items. The hate-speech items were played in individually randomized orders from the external loudspeakers of the laptop computer, using the same preset loudness level for all participants. The experimental session started with a written instruction displayed on the laptop screen. The instruction included an illustration and explanation of the 2D rating space. The instruction led over to collecting some metadata of the participants with an online questionnaire designed after [14]. After this introduction, the participants rated the hate-speech items, using an external mouse connected to the laptop computer. Each rating trial started with an automatic playback of the hate-speech item. Then, the 2D rating space was shown on the screen, in combination with the instruction to rate the heard item by clicking into the 2D space. The entire experimental procedure was programmed and run in SoSci survey [15]. Written consent to use the data was obtained per participant after the end of the experiment.

3. Results

Correlations were calculated in order to investigate whether the prosodic characteristics of the three different item classes (ORIG, HOL, IRO) influenced participants' click decisions on the x- and/or the y-axis of the 2D rating space. The x and y coordinates were correlated with those prosodic features that were found to characterize the three item classes [13] (see 2.1 above) and that are, moreover, known to be related to emotions and expressivity in a speaker's voice. The prosodic features are: f0 max, f0 min, HNR, Hammarberg index, speech rate, and formant dispersion (F1-F5), see Table 1. Spearman's rho was used for the analysis, taking into account the non-normal distribution of coordinates across raters and items.

Table 1: *Correlations of X (unacceptability) and Y (consequences) with the stimuli's prosodic characteristics.*

Prosodic feature	Dimension	*p*-value	rho
f0 max	X	0.01	0.13
	Y	0.0007	0.18
f0 min	X	0.16	-0.08
	Y	0.03	-0.12
HNR	X	0.0002	-0.21
	Y	< 0.0001	-0.30
Hammarberg index	X	0.0007	-0.18
	Y	<0.0001	-0.22
speech rate	X	0.29	-0.06
	Y	0.05	-0.11
formant dispersion (1-5)	X	0.02	0.13
	Y	< 0.0001	0.30

We found all analysed prosodic features to be significantly correlated with the rating of hate-speech items. However, there were more correlations between the y-axis and prosodic features than between the x-axis and prosodic features. For example, a lower f0 minimum made raters state with more

certainty that the hate-speech item should have consequences for the speaker. The same applied to a slower speaking rate. By contrast, there were no similar effects of a lower f0 minimum and/or a lower speaking rate on the x-axis ratings. Additionally, there were positive correlations between the f0 maximum and ratings on the x- and y-axes. The two axes were also positively correlated with the F1-F5 formant dispersion. The two voice-quality features, i.e. Hammarberg index and HNR, were both negatively correlated with ratings on the x- and y-axes.

In a further analysis, ratings along the x- and y-axes were compared between the three item classes. More specifically, the HOL and IRO items were compared against the ORIG hate-speech items, from which they were derived. Results were statistically analyzed in R by calculating linear mixed effects regression models (version 3.2.2 [16]) with subjects and items as crossed random factors, allowing for random adjustments of intercepts [17]. *P*-values of the models were calculated using the Satterthwaite approximation in the R-package lmerTest [18]. The anova()-function in R was used for the comparison of the models. Data points whose residuals were 2.5 standard deviations away from the regression line were removed and the model was refitted. In the following, values in square brackets indicate the 95% confidence interval (CI) of the estimate. Per dependent variable, 336 data points were included in the statistical analysis. Before the analysis was conducted, absolute values of coordinates were normalized by expressing them in percentages relative to the length of the respective x-/y-axis.

As to the personal (un)acceptability of a hate-speech item, the analysis yielded a main effect of feature condition (HOL: β=10.50 [2.68; 4.12], *SE*=3.57, $p<0.02$; IRO: β=-15.86 [-22.71; -9.01], *SE*=3.57, $p<0.002$). It shows that, compared to the ORIG items (81%), HOL (91%) items were rated as far more unacceptable. IRO items (64%), in contrast, were more acceptable than ORIG items (see Figure 2).

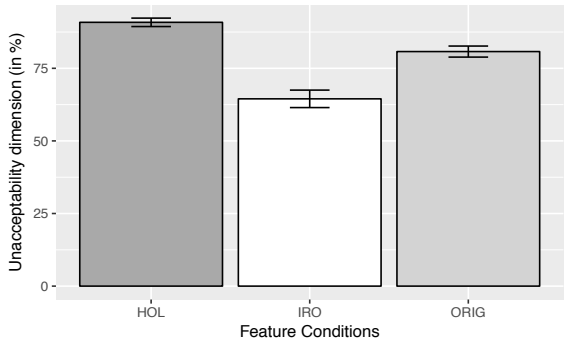

Figure 2: *Participants' normalized mean click decisions in % for HOL, IRO, ORIG stimuli along the y-axis (acceptability).*

For the sake of completeness, the statistical model was re-levelled such that ORIG and HOL ratings could be compared against IRO ratings. The model found significantly lower x-axis ratings for both ORIG items (β=-10.50 [-17.34; -3.65], *SE*=9.05, $p<0.02$) and IRO items (β=-26.36 [-33.21; -19.51], *SE*=3.57, $p<0.0001$), thus supporting that HOL items were rated most unacceptable, and showing additionally that ORIG items were rated still more unacceptable than IRO items. That is, ORIG items fell in between HOL items on the one hand and IRO items on the other.

Concerning the ratings of consequences (i.e. ratings along the y-axis), results showed a main effect of HOL (β=37.31 [24.59; 50.02], SE=6.73, p<0.0004), indicating a significantly higher consequence evaluation for HOL items (88%) as compared to ORIG items (52%). A further main effect of IRO (β=-15.32 [-28.04; -2.60], *SE*=6.73, $p<0.05$) indicated a significantly lower consequence evaluation for IRO items (38%) as compared to ORIG items. After re-levelling the model again as for the x-axis, we found significantly lower y-axis values for both ORIG (β=-37.31 [-50.02; -24.59], *SE*=6.73, $p<0.0004$) and IRO (β=-52.63 [-65.34; -39.91], *SE*= 6.73, $p<0.0001$) as compared to HOL. That is, like for the x-axis ratings, we found for the y-axis ratings that HOL items triggered the strongest and IRO the weakest consequence evaluations, with ORIG items falling in between these two item classes (see Figure 3).

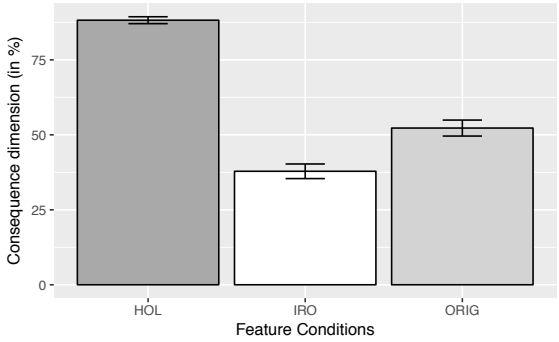

Figure 3: *Participants' normalized mean click decisions in % for HOL, IRO, ORIG stimuli along the y-axis (consequence).*

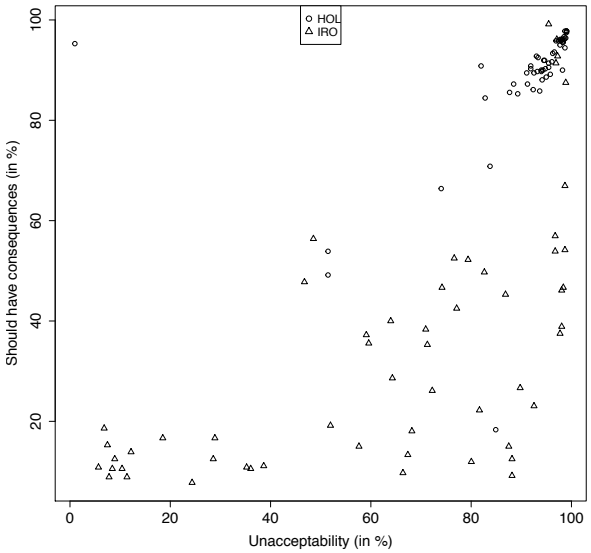

Figure 4: *Scatter plot showing participants' normalized click decisions for HOL and IRO items along the x/y dimensions.*

Additionally, Figure 4 shows how the ratings for HOL and IRO items (N=336 each) are distributed along the x- and y- axes. It can clearly be seen in this figure that HOL ratings were evaluated worst and, thus, mainly cluster in the upper right corner, whereas IRO ratings vary a lot more, particularly along the x-axis. That is, participants disagreed more on the degree to which they found the IRO items personally (un)acceptable, while they somewhat agreed that these items should rather have no consequences for the speaker. In terms of numbers, 4.8% of the IRO clicks were made in the lower left corner (i.e., in the area ranging from 0%-20%), whereas none of the HOL item clicks were made here. In contrast, only 4.5% (15 out of 336) IRO clicks were made in the upper right corner (80%-100%), whereas 42.3% of all HOL clicks were registered here; 36.3% of all IRO clicks happened in the lower half of the y-axis, as compared to only 0.9% of all HOL clicks.

4. Discussion

The present study investigated how different types of spoken hate-speech expressions are rated by German listeners along two scales: (i) participants' personal acceptability and, beyond this level of personal opinion, (ii) the perceived violation of societal norms. Our results show that HOL items receive by far the highest ratings on both scales. IRO items, by contrast, receive the lowest ratings on both scales. Thus, both items classes HOL and IRO represent – compared to the ORIG items they were derived from – two opposite poles on a two-dimensional continuum of hate speech expressions.

That IRO items yielded relatively low x/y ratings compared to both ORIG and HOL items can be well explained by their special lexical-prosodic make-up. In order to create irony, the meaning and valence of the lexical string has to clash with the meaning and valence of the coinciding prosody [19]. In the case of hate speech, this means that a negative wording coincides with a positive prosody. Besides this facilitating a higher level of rating variability, depending on what layer of spoken language participants focus on, it is typically prosody that comes off as the winner of such as a clash [20], causing lower x/y-axis ratings for IRO items. In fact, if the lower left corner (20% area, see Figure 4) with its green colored x/y-axis sections (see Figure 1) is interpreted as a hate-speech-free area, then irony is actually able to remove the hate-speech interpretation from originally clear hate-speech items. Such a conclusion would have far reaching implications for the identification and prosecution of hate speech, at least in the domain of spoken language.

In contrast, the consistently high ratings of HOL items with respect to both unacceptability and societal consequences speak for themselves. However, note that the participants were Germans. Germans have developed a particular sensitivity to negative holocaust references. It is well possible, and in fact already crystallizes in a current follow-up study, that participants with a linguistic and cultural background other than German do not make such strong ratings for HOL items.

Quantitatively, there were more correlations of the items' prosodic characteristics with the y-axis then with the x-axis. This suggests that prosody has a greater impact on how strongly hate-speech expressions violate societal norms than on how strongly they are viewed as personally (un)acceptable. The latter rating dimension is perhaps more strongly shaped by the wording of the hate-speech items. In any case, the different correlation patterns along the x- and y-axes indicate that participants were sensitive to the different meanings of the two corresponding rating scales.

The correlations between the prosody of the hate speech items on the one hand and the rating of these items on the other hand also include an interesting detail: Recall that it were lower levels of f0 minimum, HNR, Hammarberg index, and speaking rate that caused higher ratings of unacceptability and societal consequences. That is, speaking in a calm and determined way, i.e. slowly, with a low f0, a breathier voice, and in a softer, less expressive tone made hate-speech interpretations stronger, not weaker. This fact may seem counter-intuitive at first glance. However, it can probably be explained by the phenomenon of cold anger described in [21].

This explanation raises another question: Can the effect of spoken hate speech be enhanced if the respective items are realized with prosodic features that characterize hot anger? Or does the opposite apply, i.e. do items realized with hot anger (a high level of arousal) sound less serious and threatening, true to the motto barking dogs never bite? To address this question, future speakers will be asked to realize each item twice, i.e., once with a hot-anger and once with a cold-anger prosody. In this context, it is, of course, an obvious issue that our stimuli rely on only one speaker. Even though the selection of this speaker was well thought through, this fact still limits the generalization of our results. Therefore, our follow-up studies will include both more speakers, particularly female ones, and non-hate-speech baseline items from each of these speakers.

5. Conclusions

The present study focused on three main aims: First, we examined how diversely hate-speech expressions are rated along two scales of personal acceptability and the violation of societal norms. Second, we tested to what extent the prosodic parameter variation can predict hate-speech ratings along these acceptability and norm-violation scales. Third, we wanted to get an idea about the status of prosody in hate-speech ratings relative to other lexical and morpho-syntactic features.

Based on the the ratings of German native speakers, we found that hate-speech expressions containing Holocaust references yield the highest ratings on the two scales. In contrast, ironically realized hate-speech items yield the lowest ratings on both scales and may, in quite a few cases, not even be considered as instances of hate speech anymore – unlike the original (ORIG) items they were derived from and that fell in between the HOL and IRO items.

Second, even though there is a threefold gradation with respect to HOL, ORIG and IRO items on the unacceptable scale, the ORIG items are closer to the IRO than to the HOL items. This raises the question whether HOL items are an exceptional class of hate speech expressions (at least in German), while all other hate speech expressions more or less cluster with IRO and ORIG items; or whether there are also classes of hate speech expressions that create a continuum in between ORIG and HOL. Moreover, we have to test in follow-up studies to what degree the prosodic variation of hate speech items is able to influence (un)acceptability and consequence ratings and, thus, potentially interferes with the (threefold) gradation of classes of hate speech items. For example, is prosody so important that it can make an ORIG item be rated as strongly as a HOL item and vice versa?

Third, concerning the status of prosody in hate-speech ratings, it seems that prosodic parameter settings that characterize cold anger (independent of the three feature conditions introduced here) are associated with higher ratings on both the consequence dimension and the unacceptable dimension.

In a next step, our project aims at comparing participants' ratings of spoken hate speech stimuli to those of written hate speech stimuli. This will help us estimate the extent to which the spoken mode is able to weaken and/or strengthen the lexical foundation of hate speech expressions, particularly in terms of participants' physiological reactions.

6. Acknowledgements

The XPEROHS project (95-16416) is funded by the Velux Foundation. The authors would like to thank Benno Peters from Kiel University for recording the files. We thank Andrea Kleene, Nicole Baumgarten, and Klaus Geyer for their discussions and support in conducting this study.

7. References

[1] Guterres, A. (2019). United Nations Strategy and Plan of Action on Hate Speech. Taken from: *https://www.un.org/en/genocideprevention/documents/UN%20Strategy%20and%20Plan%20of%20Action%20on%20Hate%20Speech%2018%20June%20SYNOPSIS.pdf*

[2] Herz, M., & Molnár, P. (Eds.). (2012). *The content and context of hate speech: rethinking regulation and responses*. Cambridge University Press.

[3] The XPEROHS Project, main landing page: *http://xperohs.sdu.dk/*

[4] Baumgarten, N., Bick, E., Geyer, K., Lund Iversen, D., Kleene, A., Lindø, A. V., Neitsch, J., Niebuhr, O., Nielsen, R. & Petersen, E. N. (2019). Towards Balance and Boundaries in Public Discourse: Expressing and Perceiving Online Hate Speech (XPEROHS). *RASK – International journal of language and communication*.

[5] Assimakopoulos, S., Baider, F. H. & Millar, S. (2017). *Online Hate Speech in the European Union: A Discourse-Analytic Perspective*. Cham: Springer.

[6] McClay, R. (2017). Us and them: A descriptive analysis of Donald Trump's campaign speeches. Unpublished Master Thesis. University of Birmingham. URL: *https://www. birmingham. ac. uk/Docum ents/college-artslaw/cels/essays/appliedlinguistics/McClay2017. Trump-Speech-Discourse-Analaysis. pdf.*

[7] Darmstadt, A., Prinz, M., Rocholl, F. & Saal, O. (2018). Hate Speech und Fake News: Fragen und Antworten. *Amadeu Antonio Stiftung und Berliner Landeszentrale für politische Bildung, Berlin, Germany.*

[8] Niebuhr, O. & Peterson. J. M. (2011). Vorwort. *KALIPHO: Kieler Arbeiten zur Linguistik und Phonetik 1*, III-V.

[9] Baumann, S., Becker, J., Grice, M., & Mücke, D. (2007). Tonal and articulatory marking of focus in German. *Proc. 16th International Congress of Phonetic Sciences, Saarbrücken, Germany,* 1029-1032.

[10] Neitsch, J., & Niebuhr, O. (2019). Questions as prosodic configurations: How prosody and context shape the multiparametric acoustic nature of rhetorical questions in German. *Proc. 19th International Congress of Phonetic Sciences, Melbourne, Australia,* 2425-2429.

[11] Braun, B., Dehé, N., Neitsch, J., Wochner, D., & Zahner, K. (2018). The prosody of rhetorical and information-seeking questions in German. *Language and Speech 62*, 1-29.

[12] Bick, E., Geyer, K., Kleene, A (2020). "Ich habe ja nichts gegen X, aber …" - Eine korpusbasierte Untersuchung von Formulierungsmustern fremdenfeindlicher Aussagen in Sozialen Medien. In: Wachs, S., Koch-Priewe, B., Zick, A. (eds): *Hate Speech: Theoretische, empirische und anwendungsorientierte Annäherungen an eine gesellschaftliche Herausforderung*. Springer VS.

[13] Neitsch, J. & Niebuhr, O. (2019). Types of hate speech in German and their prosodic characteristics. *Proc. 1st International Seminar on the Foundations of Speech (SEFOS),* Sonderborg, Denmark*,* 85-87.

[14] Rammstedt, B., Kemper, C., Klein, M. C., Beierlein, C., & Kovaleva, A. (2013). Eine kurze Skala zur Messung der fünf Dimensionen der Persönlichkeit: big-five-inventory-10 (BFI-10). *Methoden, Daten, Analysen (mda) 7*, 233-249.

[15] Leiner, D. (2018). SoSci (Version Survey Version 2.5. 00-i). URL: *https://www.soscisurvey.de.*

[16] R Core Team. (2016). R: *A language and environment for statistical computing.* R Foundation for Statistical Computing,Vienna, Austria.

[17] Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambride University Press.

[18] Kuznetsova, A., Brockhoff, P. B., & Christensen, R.H.B. (2016). lmerTest: Tests in Linear Mixed Effects Models: R package version 2.0-32. URL: *https://CRAN.R-project.org/package=lmerTest.*

[19] Hirst, D. (2004). The phonology and phonetics of speech prosody: between acoustics and interpretation. *Proc. 2nd International Conference of Speech Prosody, Nara, Japan,* 1-7.

[20] Kohler, K. J. (2009). Patterns of prosody in the expression of the speaker and the appeal to the listener. *Frontiers in phonetics and speech science*, 287-302.

[21] Bänziger, T., & Scherer, K. R. (2005). The role of intonation in emotional expressions. *Speech Communication 46*, 252-267.