A comparative corpus study of *race* and *Rasse*

Kurt Erbach[1], Benedict Kenyah-Damptey[2], Leda Berio[3], Daniel James[2], Esther Seyffarth[2]
[1]University of Bonn; [2]Heinrich-Heine-University, Duesseldorf; [3]Ruhr-University Bochum

Abstract. Beyond the context of cross-cultural pragmatics, it has been claimed that *race* and *Rasse* are not equivalent terms. The current paper seeks to establish the first known program of study to shed light on the extent to which *race* and *Rasse* differ. Corpora of US and German newspapers are used in this paper to sample mainstream race talk in the respective countries, and top collocates are analyzed along with a selection of concordances. What is seen is that, while the historical context of Slavery and the civil rights movement in the US and Nazism in Germany do seem to be reflected in the respective corpora, it is not so clear that the underlying meaning of *race* and *Rasse* are as different as some have claimed.

Keywords. race, Rasse, cross-linguistic, American English, German

1. Introduction

Talking about race can make people uncomfortable[1] and countries differ with their national policies regarding how they deal with their past, racial atrocities[2]. Given the rich tradition of cross-cultural pragmatics (see e.g. House & Kádár 2021), it should be no surprise that discourse surrounding race can differ across cultural contexts as well. So, when Lipphard et al. (2018) claimed that the English word *race* and the German word *Rasse* ('race') are not equivalents given differences in discourse around the words, the claim was not entirely surprising. One could easily imagine that, given the different historical trajectories of the two countries, associations with the respective terms might differ in important ways. Coherently with this line of thinking, philosophers have also asked the question whether *race* and *Rasse* are equivalent terms (e.g. Ludwig 2018 i.a.). However, the claim in Lipphard et al. (2018) as well as the philosophical speculations have so far remained theoretical, and, while there is a growing body of linguistic research about how particular racialized groups are discussed in various cultural contexts (e.g. Alim et al. 2016 i.a.), to our knowledge there have not been studies on the use of the words *race* or *Rasse*, let alone comparisons between the two.

The goal of this paper is to shed light on the debate surrounding *race* and *Rasse* and the respects in which the two terms differ. While the question can extend to any other language as well, the current state of affairs has helped shape our research question. For one, a great deal of research and discussion regarding what constitutes race happens in the United States (Ludwig 2018), and for

[1]https://www.theguardian.com/commentisfree/2020/jun/18/most-of-my-white-friends-avoid-talking-about-racism-i-dont-have-that-privilege, accessed July 25, 2022.
[2] https://www.dw.com/en/world-war-ii-and-the-nazi-era-how-germany-deals-with-its-past/av-54948683, accessed July 25, 2022.

another, members of Germany's Green party recently proposed replacing the word *Rasse* in Article 3 of German Basic Law with an alternative expression on the grounds that there is no such thing as race, citing the biological notion thereof that had figured prominently in National Socialist ideology[3]. Furthermore, there has been a slew of op-ed pieces published in both US and German newspapers in the last few years about the notions of race, prompting Liphardt et al. (2018) to highlight differences between the two words. We will illuminate how the two terms differ using corpus linguistic methods.

In what follows, we first provide further background on the claims made regarding differences between *race* and *Rasse* and our subsequent research questions (Section 2.1). We will also provide some detail on two particular studies making use of corpora to investigate race talk and hate speech, from which we draw inspiration for our own empirical method (Section 2.2). We then elaborate our method in Section 3, which is a comparison of US-American English and German newspapers in order to shed light on the extent to which mainstream race talk in the different countries' dominant language differs in the ways that have been discussed. In Section 4, we see that while some of the suggested differences exist, it is not so clear that the two linguistic contexts are as entirely distinct as previous discourse might lead one to believe. In Section 5, we discuss these differences and similarities in light of our research questions, and we conclude, in Section 6, that despite some notable differences, it is not so clear that the terms are entirely distinct. We also discuss a number of challenges a cross-linguistic corpus analysis brought to light by the comparison of *race* and *Rasse* in particular.

2. Background

2.1 Discussions of race and Rasse

In a 2018 opinion piece for the New York Times, David Reich, Professor of Genetics at Harvard University, acknowledged that race is a social construct and argued that not acknowledging the possibility of substantial genetic differences between populations of humans will invite racist misuse of genetics research[4]. In a response to this piece and the subsequent discussion, Veronika Lipphardt, Professor of Science and Technology Studies and a team of other renowned German scholars at the University of Freiburg, point to a broad, interdisciplinary consensus that average genetic differences between populations sometimes do and sometimes do not correlate with self- or prescribed-membership to certain racial groups (Lipphardt et al. 2018). Moreover, they express their doubt that *race* means in English what *Rasse* means in German, which became an issue in much of the German discussion of the original piece.

In their article featured in the *Süddeusche Zeitung*, Lipphardt et al. (2018) point to several reasons why *race* and *Rasse* are different. First would be the historical perspective. *Race* in the US is tied to the history of enslaving Africans and to African Americans' subsequent struggle against racial oppression, inequality and discrimination, alongside shifts in which immigrant groups are

[3] https://www.dw.com/en/greens-call-for-race-to-be-removed-from-german-constitution/a-53733161, accessed July 25, 2022.
[4] https://www.nytimes.com/2018/03/30/opinion/race-genetics.html

identified as non-white–e.g. Jewish and Italian immigrants being possibly considered non-white early on (Gilman, 2000; Luconi, 2011). In the present day, race or ethnicity is something US residents are accustomed to self-identifying and indicating as part of demographic data in paperwork including the U.S. Census.

Quite differently, Lipphardt et al. (2018) point out that *Rasse* became entrenched in everyday German language during the rise of Naziism, and these racist ideologies could even be found in German-language textbooks until the 1990s. Lipphardt et al. (2018) also comment on how, in the US it seems there is little debate about racial categories, which reflect assignment from others and the self as opposed to biology, and can be changed based on context, while in Germany it seems that many people would take a skeptical, undecided, or curious attitude towards the question of whether or not race exists. Moreover, they note that Germany has specific conventions for talking about race, including a wide range of ways of speaking, not only because of the history, but also because of a fear of polarization and being attacked as a racist. Despite these histories and present attitudes, Lipphart et al. (2018, p. 13) argue, in both contexts there is an "unscientific and irresponsible" belief that races and ethnicities are a biological reality.

In addition to the more public oriented discussion of race, and *race* and *Rasse*, there has been relevant philosophical work in this area as well. It has been noted that much of the debate about race is US centric (Ludwig 2019), though the use of *Rasse* has received some attention. Plümecke (2014) takes a stronger position than Lipphard et al. (2018) on the relationship between *Rasse* and Germany's past, stating that *Rasse* is semantically entangled with the eugenic policy of Nazis. Ludwig (2019) goes even further and states that *Rasse* is still associated with Nazi ideology in present day Germany.

While Lipphardt et al. (2018) have made a strong case against any equivalence of *race* and *Rasse* based on historical context, given the lack of linguistic evidence we see the question of how similar *race* and *Rasse* are as an open one. While the historical contexts undoubtedly shape the meaning and use of words, we argue that it is insufficient to assume a pair of words like *race* and *Rasse* are so different based on historical evidence alone. Instead, empirical evidence is needed to support the claims made about how *race* and *Rasse* differ. Moreover, the existing discussion on *Rasse* in particular suggests that there are particular sentiments about the use of *Rasse*, and a strong or undeniable connection to the racial ideology of Nazism. Together all of this discussion leads to at least the following interrelated research questions: To what extent do the use of *race* and *Rasse* differ? To what extent can differences in sentiments be detected? And, to what extent are the use of these words explicitly tied to historical contexts?

2.2 Previous corpus based work on related issues

For the present study, we draw on two specific previous studies as models for our own approach to the comparative study of *race* and *Rasse*, namely Baker & Levon (2015) and Cabot et al. (2021). Though Baker & Levon (2015) use race talk and class talk about men in the British Press as the basis for their comparison of qualitative and quantitative corpus methods of Critical Discourse Analysis, we will focus on the quantitative methods, as this will serve as the basis for our own study. Baker & Levon (2015) approached their corpus study without any pre-formed hypotheses about what the data might reveal. They used corpus analysis software to identify the 20 strongest

collocates (measured with the Dice Coefficient) of the six different groups based on racialization and socioeconomic class (*Asian men, black men, white men, working class men, middle class men,* and *upper class men*), and concordance tables were used to group the collocates according to whether or not they convey similar ideas about masculinity. For example, collocates for *black men* included *accused, defends, prison,* and *deaths*, and after concordance analysis, *accused* and *defends* were grouped with other collocates discussing this group of men as suspected of criminal behavior while *prison* and *deaths* were grouped with other collocates discussing this group of men as convicted for criminal behavior. While the quantitative and qualitative analyses resulted in both shared and unique findings, they were both able to provide generalizations about the different ways men are categorized in the British Press. For example, black men are most frequently linked to crime in the British Press, and within the context of Critical Discourse Analysis, this representation is seen as a hegemonic discourse, reproducing and reinforcing the dominant ideologies on which these representations are based (Baker & Levon 2015).

Cabot et al. (2021) annotated a corpus of Reddit comments for the purpose of building computational models that can be used to identify populist attitudes against out-groups, though we will focus on their annotation scheme, as we will adapt it for our own study. The Reddit comments chosen for the corpus were those in response to articles from the website AllSides, and comments were filtered so that they each only discussed one particular out-group. The out-groups at the focus of this study are Immigrants, Refugees, Muslims, Jews, Liberals, and Conservatives on the basis that each are the targets of populism in the US and UK. They then annotated each comment on two scales, one regarding language towards a certain group and one regarding the emotions expressed towards the group. The language categories were Discriminatory, Critical, Neutral, and Supportive, where Discriminatory alienates or portrays the group negatively, as a threat, danger, or peril to society, and/or ridiculing the group as lesser or worthless. A Critical comment was one that was critical of the group, but not to the extent of discriminatory, Supportive comments were supportive or defensive of the group, and Neutral was none of the above. What was found is that Muslims receive the highest number of Discriminatory comments while Liberals and Conservatives recieve the largest amount of Critical comments, but low amounts of Discriminatory and Supportive comments. Refugees and Immigrants receive similarly high numbers of comments from each category, and Jews received relatively low numbers of Discriminatory and Neutral comments, the majority being Neutral. Ultimately this data was used to train computational models to identify populist language, however a fuller summary of the success of these models is outside the scope of the present paper.

3. Methodology

A corpus analysis allows us to directly compare mainstream occurrences of the term *race* in US-American news outlets to those of the term *Rasse* in German news outlets to be able to assess the specific differences between the contexts of occurrence. We selected newspaper as the source of data for the sake of comparability and accessibility. The advantage of such an approach lies in the possibility to consider the context of the occurrences and evaluate which kind of use of the word occurs in the two different socio-cultural contexts and thus shed light on the nature of race talk in Germany as compared to its US equivalent.

From the outset, there are several potential drawbacks to this approach that we set aside for resolution in future research. One drawback is the fact that both *race* and *Rasse* are homonyms, the former also referring to sporting events and political contests, and the latter also referring to breeds of animals. To understand the distribution of these words across their various meanings, and in order to not eliminate any relevant examples–e.g. talking about race in the context of sports or a political contest, or comparing groups of people to groups of animals, we did not attempt to filter out such examples. Other drawbacks stem from drawing from existing corpora of newspapers, and not being sure to have a representative sample when it comes to certain discourse involving race/Rasse and the words themselves. For example, while news outlets known to politically lean left or right were collected–e.g. The Atlantic and The Wall Street Journal–no attempt was made to determine whether the corpora themselves skew left or right, despite the ability of this to impact the results. As this is the first known corpus study on the differences between *race* and *Rasse*, we treated it as somewhat exploratory and left these and other issues to future research.

3.1 Data

Examples for usages of the German term *Rasse* were extracted from a corpus containing newspaper texts from the years 1991 through 2019. The corpus is a subcorpus of the "W" archive of written language available in DeReKo (Deutsches Referenzkorpus) (Leibniz-Institut für Deutsche Sprache 2021). Usages were clustered by frequent co-occurring words, with a context window of 5 tokens to each side within the same sentence. Collocates were lemmatized in order to group forms of the same lemma together. The strength of each collocation was determined by a function provided by the COSMAS II software (COSMAS I/II).

Examples for usages of the English term *race* were extracted from the "newspaper" portion of the Corpus of Contemporary American English (CoCA) (Davies 2008). A context window of five tokens to each side was used here as well. Here, the strength of each collocation was determined with a log-likelihood ratio function that approximates the function used in COSMAS II (Rainer Perkuhn, personal communication). The formula is given in (1). Since infrequent context words receive a relatively strong LLR score using this formula, we set a frequency threshold of at least 10 occurrences of individual words throughout all three decades under investigation here.

(1) llr = 2* (a*math.log(a)+b*math.log(b)+c*math.log(c)+d*math.log(d)
-(a+c)*math.log(a+c)-(b+d)*math.log(b+d)
+(a+b+c+d)*math.log(a+b+c+d))
where a: number of observed cooccurrences of both the target and the context word;
b: number of observed cooccurrences of the context word without the target word;
c: number of observed cooccurrences of the target word without the context word;
d: number of observed cooccurrences of neither the target word nor the context word.

For each strong context word of the German or English target word, a set of 20 random example sentences from the relevant corpus was chosen for manual annotation.

Regarding the two corpora chosen, DeReKo was chosen because it is freely available and contains a massive amount of German data, stretching as far back as 1950 and still being added to

today, and COCA was chosen because it is a well balanced corpus and accessible to the authors. While both corpora contain a variety of genres of texts, to approach something resembling the common conception of *race* and *Rasse*, and to have relatively comparable subcorpora, we limited our search to only the newspaper portions of the respective corpora. Because COCA is limited to data from 1990-2019, we restricted our search of DeReKo newspapers to the same time range, resulting in two subcorpora, the German one being 5,832,393,222 words, and the English on being 122,959,393 words.

In collecting our data, we also decided to do three separate searches for each corpus, one for each decade. The purpose of this was two-fold. First, analyzing data decade by decade would allow us to avoid any anomalies that skew the results in one way or another, such as the significant increase in discussions of race following the murder of George Floyd in 2020. Second, by aggregating enough data from each decade for comparable results, we would likely have a sum total number examples of *race* and *Rasse* in the relevant senses that would allow for comparable results despite any of the occurrences of the irrelevant senses of the two words. For each decade of each corpus, we randomly selected 20 concordances from each of the top 30 collocates. This resulted in 600 concordances per decade per sub corpus resulting in a total of 3,600 examples of *race* and *Rasse* being used (1,800 in German and 1,800 in English).

3.2 Annotation

To track the way *race* and *Rasse* are respectively used in US and German newspapers, three levels of annotation were used: notional categories, attitudes, and racism. First we followed Baker & Levon (2015) in taking a theory neutral approach and simply looked at the concordances to see what sort of notional categories might exist naturally in the data. The categories we surmised are listed in Table 2, along with a generalization about what the category title means, as well as example collocates that inspired and seemed to fall into those categories. For example the collocates *religion* and *Geschlecht* (‘Gender’ or ‘sex’ depending on the context) among others were grouped under *human kind* as a mind dependent category for sorting society. Following initial grouping, the concordences were manually inspected to regroup collocates as necessary.

Human kind	(mind dependant) category for sorting society	Religion ('religion'), Geschlecht ('gender'), Klasse ('class'), religion, sex, ethnicity
Domain	*race/Rasse* as subject of discourse	menschlich ('human'), Volk ('people'), american ('american'), politic, card, factor
Subkind	particular racialized group(s)	arisch ('aryan'), verschieden ('different'), weiß ("white")
Appearance	criterion for determining race based on appearance	Farbschlag ('color'), Hautfarbe ('complexion'), color, feature, appearance
Attribution	discussing something attributed to or assumed inherent to a race	bestimmt ('destined'), Überlegenheit ('superiority'), tough, wild
Social criterion	social criterion for determining race	Abstammung Sprache Heimat ('ancestry language home'), culture
Biological criterion	biological criterion for determining race	züchten ('breed'), Zucht ('breeding'), Abstammung ('Ancestry')

Animal	having to do with animals only	Hund ('dog'), Kaninchen ('Rabbit'), Tier ('animal')
Sports	having to do with sports only	ahead, speedway, k
Politics	having to do with politics only	senate, republican, vote
Other	cannot be grouped with others	Arm ('poor'), rat

Table 1. Notional categories of use based on collocates at first glance

Going beyond the methodology of Baker & Levon (2015), we used the same notional groups to annotate each concordance with respect to the notions discussed in the context of *race/Rasse* to come to a more qualitative understanding of how the use of the two words differ. Concordances were annotated using tandem annotation (Torres 2021 i.a.), wherein annotators review data together and reach consensus annotation decisions about which notional category best applies to the concordance. In our case our team of annotators consisted of two Germans and one US-American.

The notional categories, as formulated have room for overlap, however we sought to demarcate their use in the annotation process. *Human kind*, *subkind*, and *domain*, for example, might seem warranted every time *race* or *Rasse* is used given every use might be considered a use implying the existence of race as a category for sorting people (*human kind*) into groups (*subkind*), and that race is at least in some way the topic of conversion (*domain*). However, we attempted to be quite rigid with the use of these terms. *Human kind* was generally used alongside other human kind terms such as *gender* and *ethnicity*.

(2) Gender, race, religion, **ethnicity**[5], and sexual orientation are all critical components in the court 's most @ @ @ @ @ @ @ @ @ affirmative action to national security and gay rights.[6]

Subkind was only used when a (potentially) racialized group was mentioned (3), or racialized individuals were mentioned and their membership to a racialized group was clearly implied (4).

(3) With elected officials facing criticism over the ouster of the city 's first black female police chief, Mayor John Rowe has proposed creating a committee to" explore issues of race, **ethnicity**, equity and culture" in Portsmouth.

(4) "Look at Tiger, and Venus and Serena (Williams) with tennis. They've opened those **sport**s up to all races. Kids need to see themselves in their icons. Maybe I can help."

Domain, on the other hand, occurred when race was clearly a topic of conversation. Contrasting (3) and (4) with (2), in (3) and (4) we see that race as a topic of conversation, in (3) perhaps as part of the reason a person lost her job and in (4) as one of the motivations the quoted person has for their participation in their sport, while in (2) the topic is clearly gay rights and affirmative action, and race is only mentioned in passing as part of the larger issue.

[5] Words in bold are the collocates for the respective concordance.
[6] The series of @@@ symbols indicate a portion of the text was not readable in corpus construction.

Another area in which there is potential for overlap is in that between appearance and biological markers. While characteristics of a person's appearance such as their skin color, eye shape, etc. may be influenced by their biological parents, and any mention of such things might be taken as mentions of race as a biological category, we nevertheless distinguished the two as distinct, and included other components of appearance such as hair styles and clothing choices in with the former. (5) for example was labeled with *appearance*, but not *biological*.

(5) Nor did Armani confine the racial palette of his models to just one hue. Rather, a range of races and **age**s showed up on his runway , demonstrating that good fashion isn't confined to just vanilla.

In (5) it is also not definitely the case that any specific racialized group is overtly mentioned, so *subkind* was not used here. Conversely, we did not assume that any mention of "black" or "white" entailed a judgment of skin color, rather we chose to see them only as labels entailing a racialized group, but not necessarily a skin color (see e.g. Smedley & Smedley 2005 for social construction of race). So in (6) where skin color alone is not the topic of conversation as in (5) the overt mention of *black* was only used to justify the annotation *subkind* (*Hispanic* too) and not also *appearance*.

(6) My question to the five Supreme Court justices who **vote**d to remove race as a factor is, what should we use to set the criteria for diversity in public schools and university systems beyond GPA, test scores, extracurricular activities, oh, and legacy? Many would say that this would be enough to have a diverse student body, but it's not @ @ @ @ @ @ @ @ @ @ to upper tiers when considering this criteria, unlike black and Hispanic children, who fall in the middle to lower tiers.

For *attribution* and *social criterion*, the crucial difference is that one discusses a characteristic ascribed to a group, but is not seen as a defining characteristic, and the other is the opposite. In (7), academic achievements are attributed to Asian-American people, while in (8), a social criterion for assuming Asian Americans are a racialized group is some sort of origin in Asia (e.g. place of (parents) birth).

(7) But so, traditionally, have Asian-Americans--and unlike the case with blacks and Hispanics, their race appears to be a **factor** in explaining why @ @ @ @ @ @ @ @ @ their academic achievements.

(8) Nicole Ochi, an attorney with Asian Americans Advancing Justice - Los Angeles, said Chinese language socialmedia platforms such as WeChat have stirred opposition against affirmative action. She said "flat-out lies" have been posted, such as assertions that half @ @ @ @ @ @ @ @ @ @ action is brought back in such states as California, which banned public institutions from discriminating on the basis of race, sex or **ethnicity** with the passage of Proposition 209 in 1996.

In these ways, we demarcated the use of the notional categories as annotation labels during manual inspection of the concordances.

The second level of annotation was designed to develop an understanding about the extent to which attitudes surrounding the use of *race* and *Rasse* differ, we built on the annotation categories of Cabot et al. (2021). The first difference was that we added the category *passive supportive* to be an intermediate between *supportive* and *neutral* to parallel Cabot et al. (2021)'s use of *critical* as an intermediate between *discriminatory* and *neutral*. Also, we annotated both author attitudes and any reported attitudes given the context of newspapers is such that *race* or *Rasse* might be occurring in a quote rather than the author's own words. Lastly, because multiple attitudes were reported in many instances, we added the categories *compatible-supportive*, *compatible-discriminatory*, and *opposing* to capture the instances where multiple attitudes are reported, and those attitudes are of at least two in the set {*neutral, passive supportive, supportive*}, {*neutral, critical, discriminatory*}, or {*passive supportive/supportive, critical/discriminatory*} respectively. Some examples are given in (9).

(9) a. (supportive) In den Jahren 2000 und 2002 verabschiedete die EU drei Richtlinien, die in Europa für Gleichheit zwischen den Rassen und **Geschlecht**ern sorgen sollten.
'In 2000 and 2002, the EU passed three directives that were supposed to ensure racial and gender equality in Europe.'

b. (passive-supportive) Die EG-Richtlinie verbietet nur im Arbeitsrecht eine Unterscheidung aufgrund Rasse, ethnischer Herkunft, **Religion**, Weltanschauung, Alter, Behinderung und sexueller Identität.
'The EC directive only prohibits discrimination on the basis of race, ethnic origin, religion, ideology, age, disability and sexual identity in labor law.'

c. (neutral) Der Vorsitzende des amerikanischen Leichtathletik-Verbandes meint damit den Aufstieg zu einer Persönlichkeit, deren enorme Anziehungskraft quer durch die **Geschlecht**er, Altersgruppen, Rassen und Nationalitäten geht.
'The chairman of the American Athletic Association is referring to the rise to prominence of a personality whose enormous appeal cuts across genders, ages, races and nationalities.'

d. (critical) Das Blog "Perlen aus Freital" sammelt täglich neue Hassbotschaften, und etliche Nutzer zeigen besonders hetzerische Kommentatoren an oder setzen den Arbeitgeber in Kenntnis. Eigentlich verspricht Facebook, dass "sämtliche Hassbotschaften", die Personen "aufgrund von Rasse, Ethnizität, nationaler **Herkunft**" angreifen, sofort entfernt würden - doch das sei ein leeres Versprechen'so der Vorwurf vieler Nutzer.
'The blog "Perlen aus Freital" (Pearls from Freital) collects new hate messages every day, and quite a few users report particularly inflammatory commentators or inform their employer. Facebook actually promises that "all hate messages" that attack people "on the basis of race, ethnicity, national origin" will be removed immediately - but that is an empty promise according to the reproach of many users.'

e. (discriminatory) Die Klageschrift, die James Damore und ein weiterer ehemaliger Google-Mitarbeiter bei einem kalifornischen Gericht eingereicht haben, enthält viele befremdliche bis lustige Sätze. Mir persönlich gefällt dieser besonders gut: "Googles offene Feindseligkeit gegenüber konservativem Gedankengut geht Hand in Hand mit unfairer

Diskriminierung auf der Grundlage von Rasse und **Geschlecht**, was das Gesetz verbietet". Gemeint ist Damores eigene "Rasse" - er ist weiß - und sein eigenes Geschlecht.
'The complaint filed by James Damore and another former Google employee in a California court contains many strange to funny sentences. Personally, I particularly like this one: "Google's open hostility to conservative thought goes hand in hand with unfair discrimination based on race and gender, which the law prohibits." Meaning Damore's own "race" - he's white - and his own gender.'

In (9a), we see a supportive attitude in the report of three EU directives designed to ensure racial equality in Europe. In contrast, (9b) is similar albiet passive-supportive given it only prohibits racial discrimination, meaning the goal is not to have people be euqal, rather it is for people to just not be overtly discriminated against in terms of race. We see this as passively supportive because it does not seek to uplift the oppressed as much as it aims to prevent further oppression. By comparison, (9c) is totally neutral saying nothing positive nor negative about any particular oppressed racialized group.

While certain attitudes like those of the Nazis are uncontroversially racist and simple to annotate, others present somewhat more of a challenge. (9d) for example contains mention of hate-speech but not the hate-speech itself. For this reason we annotated the reported attitude as critical, given it could not be judged for its degree of negativity. Other instances, as in (9e) involve contestation over whether some occurrence even counts as an instance of racism. In this case, the reported speech involves the claim that the speaker is discriminated against *because he is white*—a claim that the author of the quoted passage finds particularly "funny" [lustig], thereby suggesting that they dispute it. Plausibly, contesting the claim that the speaker has experienced racist discrimination (because he is white) rests on the widespread idea that racism is, above all, a matter of structural oppression, rather than individual discrimination (see e.g. Bonilla-Silva 2006; Feagin 2006; Roig 2017; Urguidez 2020). Thus, (9e) amounts to a metalinguistic (as opposed to a factual) dispute over *which concept of racism to use* (see, e.g., Plunkett & Sundell 2013; Plunkett 2015; Urquidez 2020).

The third level of annotation was developed in order to better track the extent to which the use of *race* and *Rasse* differ when it comes to reference to either present or past instances of racism. In other words while some concordances like that in (10) made direct reference to a past instance of racism, in this case the administrations of Harvard and MIT excluding Asian people after a certain number had already been admitted, others like that in (9d) do not make direct reference to an instance of racism but do so indirectly, in reports of hate speech, and lastly others like that in (9c) make no reference to racism, direct or indirect.

(10) The study's author, Althea Nagai, looked at acceptance rates at the three schools and found that the two that use race and **ethnicity** as factors in admission, Harvard and MIT, appear to cap Asian acceptance rates, much as rates of acceptance for Jews were limited by elite schools in early eras.

4. Results

This section presents the results of the corpus study after manually inspecting each of the 3,600 concordances collected. First the collocates will be represented in groups connected to their use. Then the extent to which the notions surrounding *race* and *Rasse* were judged to occur by our annotators will be presented, followed by the attitudes of authors and those reported. Lastly, the extent to which an instance of racism is overtly, indirectly, or not mentioned at all will be reported, and the results will be discussed with respect to the research questions and particular challenges in the subsequent section.

After inspecting the concordances manually, we did not need any new notional categories, though we did reorganize the distribution of the collocates as in Table 2, which omits those pertaining to animals (from the German data) and sports and politics (from the English data) for reasons of space..

Human kind	category for sorting society	Abstammung ('Ancestry'), all- ('all'), Ethnie ('ethnicity'), ethnisch ('ethnic'), Geschlecht ('Gender'), Herkunft ('origin'), Klasse ('class'), Nationalität ('nationality'), oder ('or'), Religion ('religion'), Volk ('people'), age, color, culture, ethnicity, religion, sex
Domain	race/Rasse as subject of discourse	menschlich ('human'), politic
Subkind	particular (racialized) group(s)	arisch ('aryan'), jüdische Vernichtung ('jewish extermination'), nordisch ('nordic'), unterschiedlich ('different'), weiß ('white'), zwischen ('between')
Appearance	criterion for determining race based on appearance	Hautfarbe ('complexion'), appearance
Attribution	discussing something attributed to or assumed inherent to a race	aussterben bedroht ('endangered'), minderwertig ('inferior'), Überlegenheit ('superiority')
Social criterion	social criterion for determining race	Abstammung Sprache Heimat ('ancestry language home'), Kultur ('culture'), Nation ('nation')
Biological criterion	biological criterion for determining race	

Table 2. Notional category membership based on collocate use

While most collocates were aptly sorted pre-inspection several were not. *All-* (‘all’) and *Volk* (‘people’) were assumed to occur in the context of discussions of distinct subkinds, however, upon inspection, specific groups were not overly mentioned in most contexts in which these collocates occurred rather they were more strongly associated with being simple associated with sorting people in different ways, and thus recategorized under human kind. *Herkunft* (‘origin’) and *ethnisch* ('ethnic'), were assumed to be a social criterion for determining membership to a group considering, for example the sometimes interchangeability of *Black* and *African American* in the case of *Herkunft* (‘origin’), however, they were more commonly seen as other means of sorting people in general as well. *Abstammung* (‘ancestry’) was grouped with other collocates that seemed to suggest focusing on biological criteria, and while this was found to be the case in several concordances, the

majority were not as overly biologically focused and the term was just used as a means of sorting people into categories and it was thus recategorized as human kind.

Nation (‘nation’), like *nationality* was assumed to be a way of sorting people into categories, thus in human kind, and while this was largely the case these categories were tied to distinct racialized groups as in, and so *Nation* (‘nation’) was recategorized under social criterion rather than human kind. *Jüdische Vernichtung* ('jewish extermination') was originally set aside given it was assumed to refer to an event, the genocide of Jewish people carried out by Nazi regime, though it was recategorized as pertaining to subkind given the discussion of Jewish people as well as a racialized group.

While *bestimmt* (‘destined’) and *gefärlich* (‘dangerous’) were assumed to be attributive to a particular racialized group in discussions of Naziism, and though they did occur a couple of times in this manner, they were seen to be used primarily for animals and recategorized as such. Similarly, *american* (‘american’) was assumed to be domain based on discussion of race in the context of the US, *Farbschlag* (“color”) was assumed to be best categorized as appearance, and *verschieden* (‘different’) was thought to be prominent in discussions of subkinds though these were recategorized as animal based on their use primarily in such contexts. It is also worth noting that *aussterben bedroht* ('endangered') was accurately assumed to pertain primarily to animal usage, though it was also seen in a few instances to pertain to humanity as a whole.

With respect to the English data, the vast majority of the collocates pertained to sports and no recategorization was necessary though it is worth noting that several relevant examples of *race* were found during inspection of these collocates. Particularly frequent, though not a majority were instances where *sport* was a collocate in discussion of race in the relevant sense (4). While *color* was assumed to be used in discussion of appearance in the context of race, it was seen to be used mostly as a category for sorting people alongside race (19) and thus recategorized as human kind. Similarly, *culture* was assumed to occur as a means of discussing race in terms of culture, though it most frequently occurred as a category for sorting people into alongside *race*, (20).

Alongside the collocates that were re-categorized in another group pertaining to the relevant sense of race, many collocates were first thought to pertain to the relevant sense but turned out to pertain to sports or politics. *Card* was thought to be domain because of talk about “playing the race card” meaning to bring up race to make a poitn, but the majority were found to pertain to sports in the phrase “wild card race”. *Feature* was assumed to be relevant in discussions of appearance and *factor* assumed to be relevant in discussions of race in general and *wild* an attribution to a racialized group, yet these three collocates turned out to be almost exclusively used in the context of sports, such as features or factors of athletic races, and wild card races in particular. *Tough* was thought to be an attribution to a racialized group but ended up being most common in discussing tough political races. *Enter, stage,* and *Virgina* were thought to be more common with examples like entering sports races, stage races, and athletic races in Virginia but were slightly more commonly relevant in examples like entering a political race, a stage of a political race, or a political race in Virginia. Of the 3,600 concordances inspected, 2,655 were found to be irrelevant given they did not discuss race or Rasse in the relevant senses, meaning 945 examples (687 German, 258 English) constitute the basis of the results in what follows.

Looking at the re-categorization of the collocates, the most drastic change is that the category of biological criterion has lost all colocates. Attribution, social criterion, and subkind and

social criterion have likewise lost all English collocates, the majority of which went to sports and human kind. Several German collocates were likewise lost from these notional groups and added to animal.

Given the collocates could have been categorized under multiple notional categories, and given multiple notions related to *race* and *Rasse* can occur simultaneously, we also tagged each concordance according to which notions appeared in the co-text. The annotations were aggregated in terms of their percentage of occurrence across the relevant examples in each decade, and the averages across each decade per language are displayed below in Figure 1.

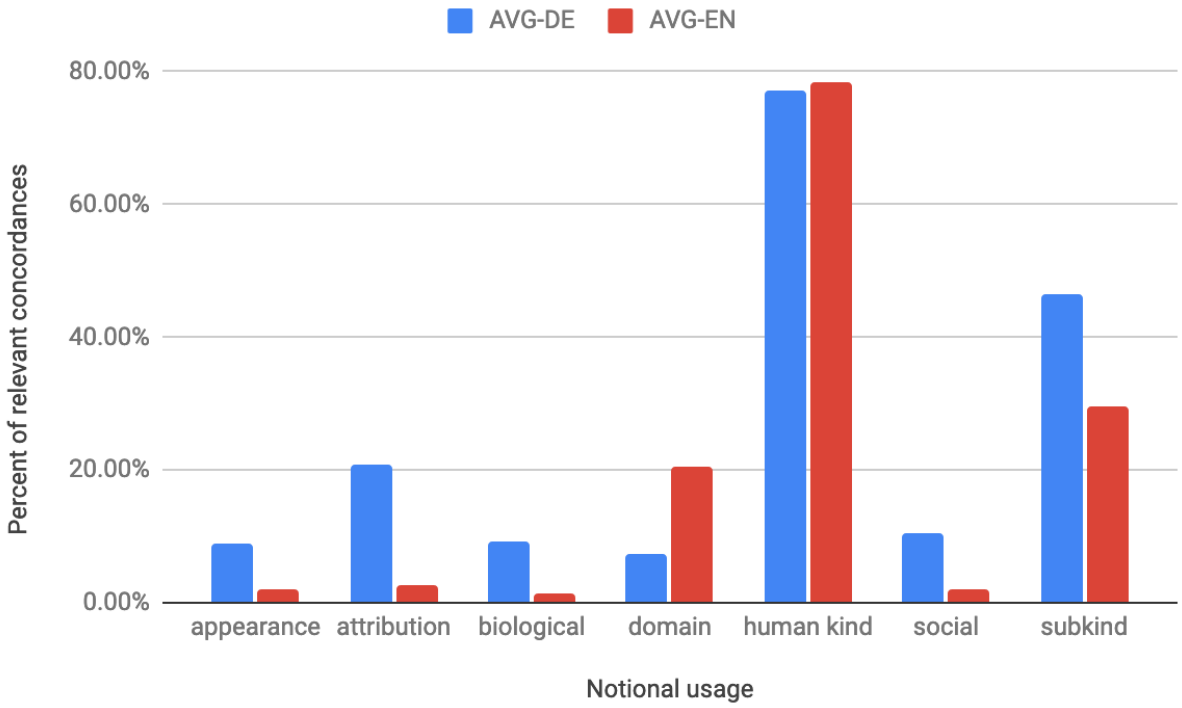

Figure 1. Notional usage in percentage of concordances, average across 90s, 00s, and 10s

As seen in Figure 1, the vast majority of instances of *race* and *Rasse* are in the context of the human kind notion (77.20% DE, 78.35% EN), where society is sorted into groups, as is likewise seen in the large number of collocates–e.g. *religion*, *nationality*–that do likewise. The second most frequently occurring notion is that of subkinds (46.39% DE, 29.64% EN) where a group name is overly mentioned, or in some instances specific individuals were mentioned as exemplars of an otherwise unnamed racialized group as occurs with Tiger Woods and Venus and Serena Williams in (4). While second highest in both contexts, the higher frequency in the German context is due to the high frequency of mentions of *arisch* ('arian'), *weiß* ('white'), and *nordisch* ('nordic') often in discussion of Naziism.

The remaining notions all occurred with much less frequency then the other two. As with subkind, most notions occurred more frequently in German than in English: appearance (8.98% DE, 1.91% EN), attribution (20.87% DE, 2.69% EN), biological criterion (9.21% DE, 1.28% EN), and

social criterion (10.45% DE, 2.03% EN). The one standout notion is that domain (7.40%DE, 20.45%EN) occurred more frequently in English than in German.

For the sentiment analysis, recall that each concordance was annotated with respect to the author's attitude towards racialized groups, as well as any attitudes reported. When it comes to author attitudes, as shown in Figure 2, neutral attitudes were most frequent by far in both the German and American corpora (80.49%DE, 74.97%EN). While all other attitudes were much less frequent than neutral, some are somewhat noticeable differences: discriminatory (1.23% DE, 2.99% EN) and critical (3.74%DE, 4.87%EN) were both lower in German than English. While passive supportive (9.43%DE, 6.79%EN) was higher in German than English, supportive (1.54%DE, 7.86%EN) was far lower in German than English.

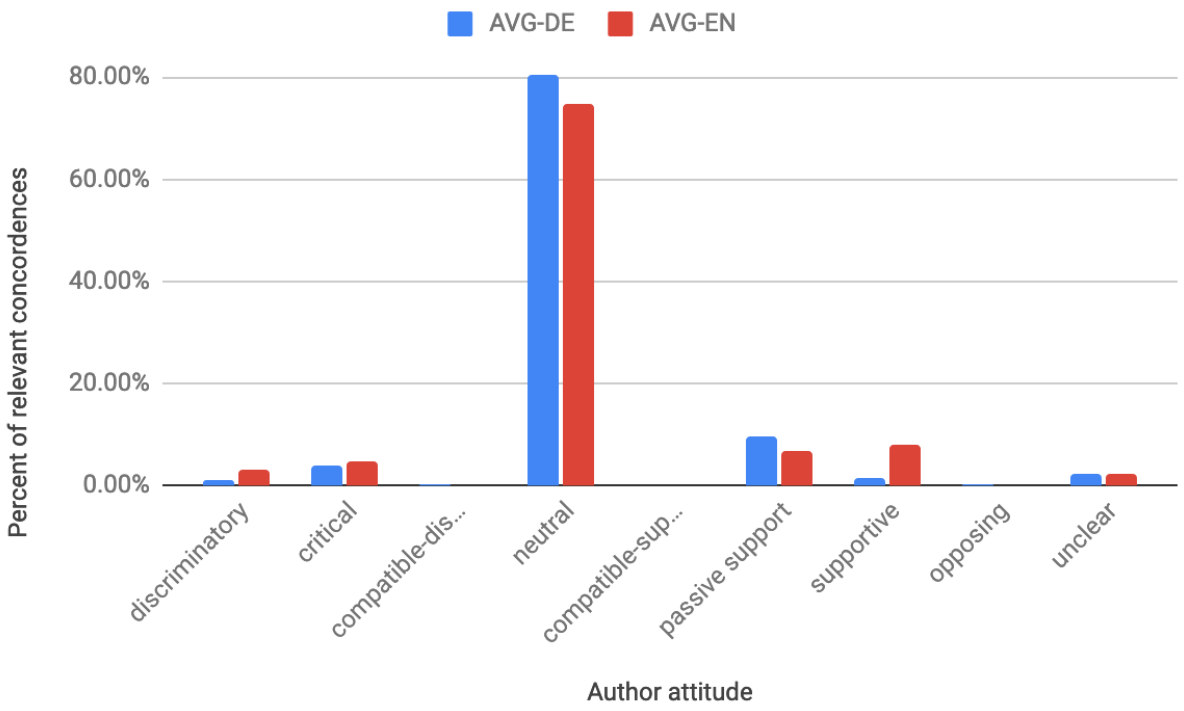

Figure 2. Author attitude in percentage of concordances, average across 90s, 00s, and 10s

Reported attitudes are more evenly distributed across the categories, as shown in Figure 3, though some categories occur very infrequently and at least one quite frequently by comparison. Discriminatory attitudes (38.46%DE, 25.48%EN) were the most commonly reported in both corpora, though quite higher in German than English. Though much smaller by comparison, the opposite is seen with critical attitudes (2.39%DE, 7.20%EN) German being far lower than English, but not compatible-discriminatory attitudes (4.78%DE, 2.55%EN). Together, these negative attitudes (45.64%DE, 35.23%EN) are more common in German than English. By comparison, the cumulative positive attitudes (21.99%DE, 20.25%EN) were nearly identical across the two corpora, though much less frequent than the total negative attitudes. Supportive attitudes (13.92%DE, 11.16%EN) were slightly more common in German than English, same with compatible-supportive

attitudes (1.17%DE, 0.29%EN), though the opposite was seen with passive supportive attitudes (6.90%DE, 8.79%EN). Neutral attitudes (15.07%DE, 19.58%EN) were fairly common and slightly higher in English, as were opposing attitudes (14.84%DE, 18.31%EN).

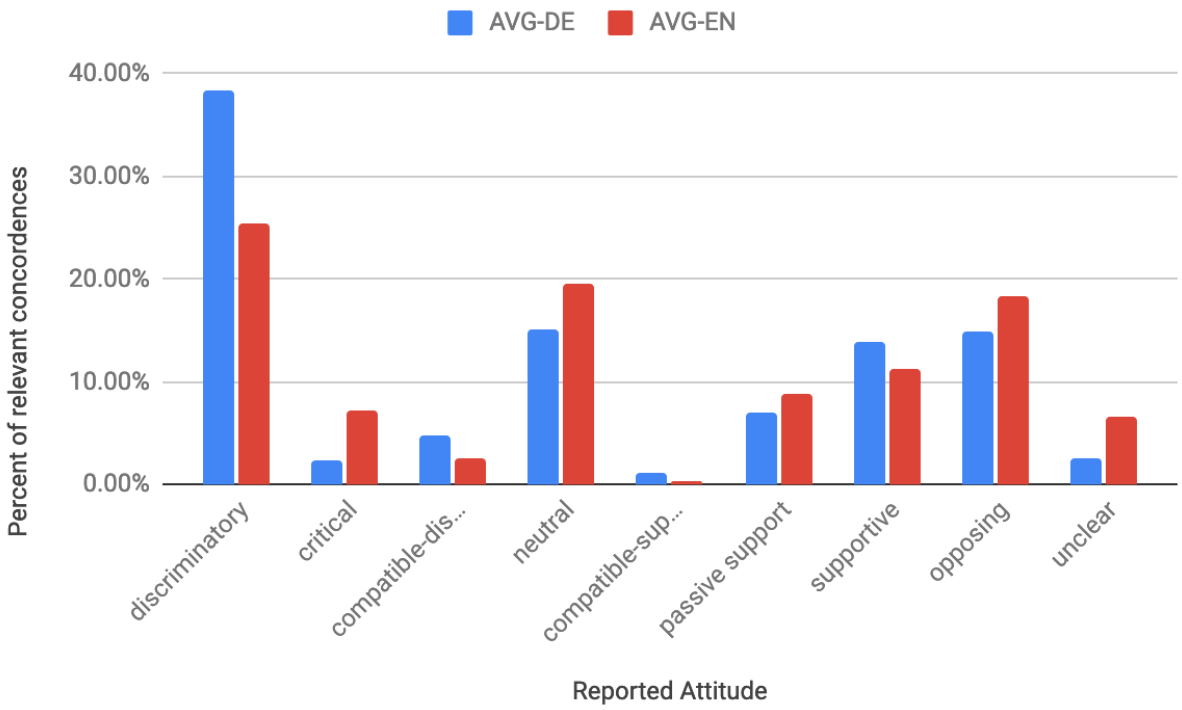

Figure 3. Reported attitude in percentage of concordances, average across 90s, 00s, and 10s

Whether reported instances of racism occured in the concordance also generally varied across the corpora. They occurred in more than half of the German examples, very often being mentions of Nazi related racism, and less frequently in English in various forms from segregation in cities to biased admissions practices in universities (55.05%DE, 36.75%EN). Indirect mentions were fairly common and even across both (23.37%DE, 25.50%EN), and instances of *Rasse* without an overt or indirect mention of racism were somewhat uncommon in German, often in reference to German Basic Law, but more common in English, sometimes in reference to similar, anti-discrimination laws (19.28%DE, 32.79%EN). In general, the German concordances generally mentioned an instance of racism, while the English concordances were somewhat evenly distributed across the three categories.

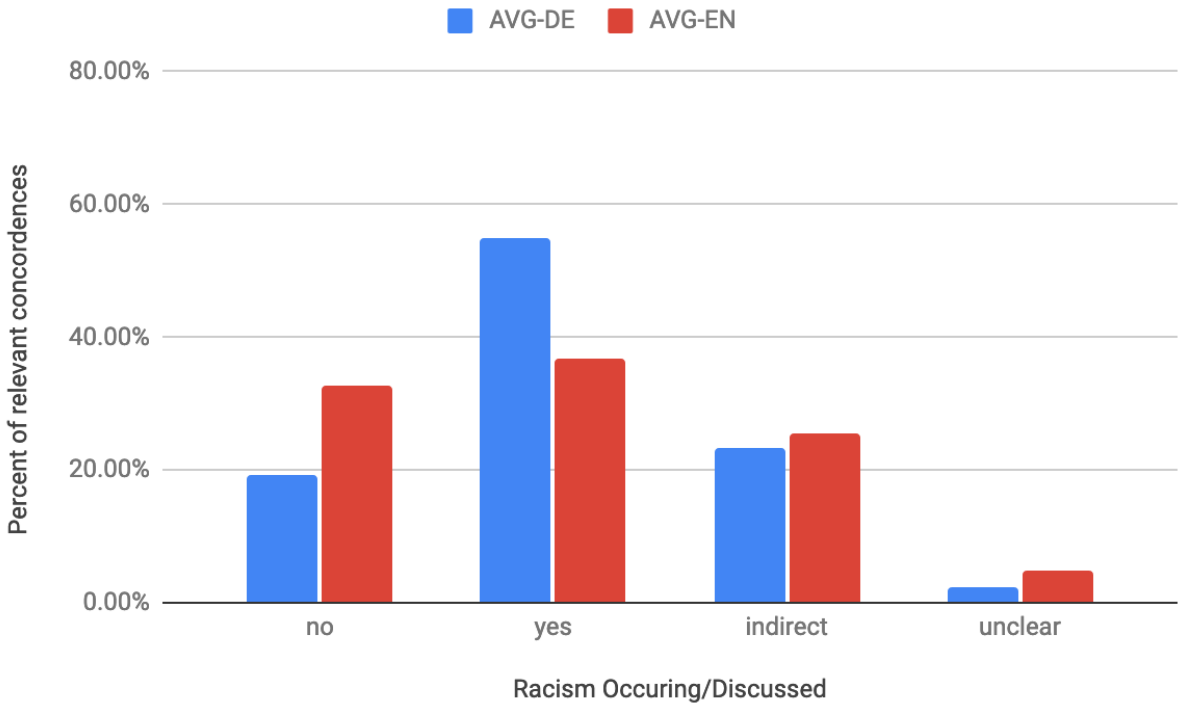

Figure 4. Instances of racism in percentage of concordances, average across 90s, 00s, and 10s

5. Discussion

5.1 Answering research questions:

To what extent do the use of *race* and *Rasse* differ? With the exception of the vast differences in homonymy–race also referring to political and atheltic contests as well as a particular manner of movement and Rasse also referring to animal breeds–the differences between *race* and *Rasse* as they refer to racialized groups of people are smaller than one might expect given Lipphardt et al. (2018) claim that race and Rasse do not mean the same thing given their distinct histories in their respective contexts. There are some undeniable differences in co-text as seen in the collocates, such as Rasse being frequently discussed in the context of Nazism, including more attribution terms like *(vom) Aussterben bedroht* ('endangered'), *minderwertig* ('inferior'), and *Überlegenheit* ('superiority'), more subkind terms like *arisch* ("aryan"), *jüdisch Vernichtung* ('jewish extermination'), *nordisch* ('nordic'), *unterschiedlich* ('different'), *weiß* ("white"), and *zwischen* ('between'), and more social criterian terms like *Abstammung Sprache Heimat* ('ancestry language home'), *Kultur* ("culture"), and *Nation* ('nation'). Nevertheless, the two languages have similarly large (proportionally) numbers of human kind terms like *Religion* ("religion"), *Volk* ('people'), *age*, *color*, etc. So despite the larger number of distinct collocates in German, which are often used in the context of discussion of Nazism, both langauges use their respective terms in a similar way, namely

as a category for sorting people, as made evident in the numerous examples discussing (law(suit)s against) discrimination. In this way, this use of the respective terms seem to at least acknowledge that some people might believe race to exist in some way, and that it may end up being the basis of discrimination despite the laws against it. Another similarity is that no collocates ended up being categorized as most frequently expressing biological criterion for characterizing race/Rasse, and in this way one might think that neither US-Americans nor Germans conceptualize race/Rasse in this way, however the notional use analysis suggests otherwise.

While *race* has less diverse collocates than Rasse, the notional use annotation serves to mitigate a number of these differences. The collocate analysis would make it seem that English contains no discussion of attributions, social criterion, or subkinds, however, the notional use analysis shows that these notions do occur in the English data as well, albeit with a lesser frequency than in German. Having one collocate each in the domain category and the appearance category would suggest relatively equal frequency of these notions in discussion of race/Rasse. However the notional analysis shows that notions of race/Rasse tied to appearance are more frequent in German than English and those tied to domain are less frequent in German than English. Finally, while no collocates were most frequently tied to discussions of biological criteria for race in either subcorpus, notions of biological criteria were seen in both subcorpora, far more in the German data than in the US-American data. Unsurprisingly, human kind notions dominate the data in both contexts.

All together, we take the collocate and notional use data to suggest that there are indeed some differences in how *race* and *Rasse* are used as suggested by Lipphardt et al. (2018), however there are some similarities as well. Supporting Lipphardt et al. (2018) are the German collocates frequently tied to discussions of Nazism, and the notional usage analysis where *race* seems to be more frequently a topic of discusison given the frequency in the domain category. The overall distribution of both collocates and notional usage support these differences as well. However, what undermines the claims and data showing differences is the fact that the vast majority of uses of *race*/*Rasse* as a human kind term, alongside other words doing the same, in discussions of discrimination. Moreover, while the frequencies across the other categories differ, the fact that each category is present to some degree in the notional usage of each term suggests that the underlying concepts have the same sort of attributes, even if they are discussed to varying degrees and in varying ways across the two contexts. In other words, the overt discussions that include *race*/*Rasse* are prone to differences, but there are undeniable similarities that might be taken to undermine the doubt expressed by Lipphardt et al. (2018) that *race* and *Rasse* mean the same thing, depending on one's theory of meaning.

To what extent can differences in sentiments be detected? Similar to the results of annotating the notional use of race/Rasse, annotating the author and reported attitudes showed broad-scale similarities and small scale differences. Starting with author attitudes, we find the full sepectrum of attitudes from discriminatory to supportive across both the German and US-American data. Perhaps unsurprisingly, assuming the press is supposed to be unbiased and only report different sides of an issue, the author attitudes are largely neutral, and positive and negative attitudes make up a small fraction of the author attitudes in both contexts. While positive author attitudes are, collectively, fairly similar in frequency (10.97%DE, 14.65%EN) across the contexts, supportive author attitudes are much less frequent in the German context than in the US-American

context (1.54%DE, 7.86%EN). This difference might be taken as indicative of the purported discomfort that Germans are said to have regarding Rasse--the idea being that their discomfort with the concept given the strong association with Naziism prevents them from using the concept to uplift oppressed racialized groups--while US-Americans have, to at least some degree, embraced the concept as a means of motivating such ameliorative behavior. At the same time, negative author attitudes are, both collectively and per category, less frequent in German than English (discriminatory: 1.23%DE, 2.99%EN; critical: 3.74%DE, 4.87%EN, total: 4.97%DE, 7.86%EN), so while negative author attudes towards minoritized racialized groups are by far less frequent, they are still present in both contexts. In terms of author attitudes, it is therefore not clear that Germans have a particularly distinct way of talking about Rasse than do US-Americans about race, contra Lipphardt et al. (2018)

When it comes to reported attitudes we again saw largely similar patterns across the contexts with sometimes smaller and sometimes larger differences. While we did see more discriminatory attitudes in the German context than the US-American one (38.46% DE, 25.48%EN), we saw the reverse with critical attitudes (2.39%DE, 7.20%EN). Together with the compatible-discriminatory attitudes (4.78%DE, 2.55%EN), the cumulative negative attidues do differ with there being more in the German context than the US-American one (45.64%DE, 35.23%EN). Given many of these negative attitudes are reports of Nazi ideologies, the difference might be taken to support the claim of Lipphardt et al. (2018) and others about the strong ties of the word *Rasse* to Nazism.

However, the difference between the German context and the US-American one is not so drastically different that it would suggest there are entirely different attitudes about Rasse and race. Quite oppositely, the fairly similar distribution of attitudes across the contexts suggest the two terms are more similar than has been claimed. Especially on the positive side, each category differs by less than three percent, and the difference between opposing views is not much larger. Given positive attitudes are less frequent than negative ones in German, we might have an explanation for the claim that Rasse has an overall negative connotation, however since the difference is not that distinct from what is seen in English, what is left unanswered is why the same negative connotation is not also had in English, and for this, we look towards the instances of racism reported.

To what extent are the use of these words explicitly tied to historical contexts? The final category of annotation looked at the extent to which instances of racism were reporte/occured in the concordance, and we found that they occured overtly far more frequently in the German than US-American context (55.05%DE, 36.75%EN), they occured indirectly a nearly equal amount (23.37%DE, 25.50%US). Coupled with the more discriminatory attitudes towards minoritized racialized groups, the larger number of reported/occuring instances of racism in the German data further support the idea that Rasse is closely intertwined with Nazism and is something that can make the average German uncomfortable. Given race occurs more frequently outside explicit or implicit (mentions of) racism in the US context (19.28%DE, 32.79%EN), this too supports the idea that US-Americans use the word more freely without concern of being accused of racism, unlike Germans do according to Lipphardt et al. (2018).

In general, concerning the various associations that have been made between the word Rasse and Nazism, while the present study cannot prove that there are instances where the use of the word is not tied to Nazism, there certainly appear to be some. Instances where there are neutral

or positive attitudes towards minoritized racialized groups and no instance of racism is overtly or indirectly mentioned are those in which Rasse is less expicitly tied to Nazism. Given these results, it's at least the case that Rasse is not overtly tied to Nazism or semantically entangled with the mass execution of millions of Jewish people, hundreds of thousands of Polish and Romani people, and tens of thousands of gay men, and more than a thousand Jehovas Witnesses. Whether there is a deeper connection will have to be the work for other empirical approaches.

5.2 What do we talk about when we talk about *race* and *Rasse*?

While the collocate and concordance analysis have shed some light on the research questions regarding the differences in the use of *race* and *Rasse*, there are still many details about these words and their use left to be uncovered. In this subsection, we will review some of the peculiarities about these words that struck us during the tandem inspection and annotation of the concordances.

It seems that, so long as *race* and *Rasse* are used, they seem to assert they can be used to sort people into categories of some nature. In both the German and US-American data, one can find passive-supportive and neutral examples where race seems to be assumed to exist and can be used to sort people as a human kind like sex or religion, whether or not they should be, and a subkind may or may not be mentioned. These stand in contrast to the opinion sometimes seen in German that *Rasse* is a purely biological notion that does not exist in humans.

As for how *race* and *Rasse* are seen, there is definitely a mix of views. On the one hand we see overtly biological views such as those where the idea of drugs reacting unfavorably with people of a certain race only makes sense if there is a biological component. While US researchers state that race is a social construct and is thus based on social criteria for membership, it is not always clearly the case that race is viewed in this way. Notice, however, that these discussions never seem to include debate about biological factors. Rather, the instances where biological markers are mentioned, they seem to at least imply that some people still assume race has biological manifestations such as in DNA or in the body's reaction to certain medications or is something that can presumably be breeded out by allowing immigrants into Germany. At the same time, in both corpora there are examples of people explicitly stating that there is data against a biological view of race.

At the same time we see other views wherein race and ethnicity are seemingly intertwined in some way. In one German concordance, it was discussed how the editors of a book replaced all instances of *Rasse* with *Ethnie* ('ethnicity') and that the author had no problem with this, which at least suggests the former can be subsumed by the latter. The ability to replace *Rasse* with *Ethnie* ('ethnicity') might otherwise entail that both race and ethnicity are biological and social to varying degrees, making the two terms sufficiently synonymous.

Taken together, the discussions involving *race* and *Rasse* that the concept of race is alive and well in both contexts. In some cases it seems taken for granted that race exists in some way, others continue to assert the existence of a biological view despite the several well-researched reasons to doubt such a view (see e.g. Mukherkee, 2016). So while scientific consensus is such that racialized groups cannot be distinguished according to genetics, discussions of biology and race do still intersect in such a way that suggest biological views of race are somewhat commonplace.

Recall that Lipphardt et al. (2018) stated that Germans largely avoid mention of race so as to not be deemed a racist given the strong association between *Rasse* and Nazism. This claim is made evident by examples in the corpora that state things like '[the far right] babble about the hierarchy of races', or that people have a "fear of getting entangled in a touchy dialogue on race and ethnicity." Moreover, it seems that discussions of race and racism are sometimes dismissed as emotional blackmail of white people as seen in several examples where people of minoritized groups are accused of "playing the race card" in the English data. This seems to be a further strategy of avoidance as well as further oppression: by dismissing discussions of race, one can attempt to ignore the systemic injustices and inequalities that continue to harm historically oppressed groups.

Of course, as the data shows, mention of race is not always avoided, given there are plenty of instances in the corpus. As a topic under discussion, items seem to be more focused on race relations or treatment of a certain racialized group rather than discussing what it means for something to be a race, though, note there was one example asserting "Hispanic is an ethnicity, not a race", so there is some amount of discussion about what constitutes a race. Importantly, there is no positive evidence that discussions of race/Rasse avoided in either the German or US-American context because race is seen as a biological category that does not exist, rather, race is very clearly assumed to exist in some way or at least to some people given its mention, and people seem to avoid it for other reasons.

6. Conclusion

In conclusion, we have provided evidence for the differences between *race* and *Rasse* alleged by Lipphardt et al. (2018) via a comparative corpus study of US and German newspapers. They claimed that German Rasse is not equivalent to English race based on the historical context and impressions about how Germans and US-Americans behave in the context of the use of the respective words, and we have shown that the historical context discussed, namely strong ties to Nazism in Germany and ties to demographic identification in the US are indeed borne out. We have also shown evidence of the notion that Germans seem to associate discussions of race with Neo-nazis and far right political groups, and for this reason might avoid such discussions in general, and we have also shown that discussions of race are also avoided by certain people in the US, for fear of offending someone or because bringing up race can be deemed emotional blackmail. This avoidance in the US-context is one similarity not discussed by Lipphardt et al. (2018). Another similarity seen that was not anticipated is that race seems to be discussed in the same variety of ways (appearance, attribution, biological, domain, human kind, social, subkind) across the two contexts, albeit to different degrees, which does support a distinction between the two terms. In general, while distinctions can be made between the two terms, it is not clear that they are as different as some might believe.

Stronger claims about the use of race and Rasse are somewhat limited by the nature of the data. For one, the largely quantitative work does not include room for lengthy discussion about what may or may not be meant by the respective terms in each use in our data set. Moreover, the data sets themselves might not be as revealing as they otherwise could be, because of the attempts to be as theory neutral as possible, and to have as comparable data sets as possible. Refining search criteria might be one way to have a richer data set, though this would require limiting the respective searches in the German and English corpora in certain ways. For example, exploratory work on the

English side has shown that limiting a search to "of|on|about|by|other race_nn" can yield 11,959 instances of *race*, close to 10% of the total 112,507. Inspecting 10 random concordances from this subset yielded only one instance of *race* in a non-relevant sense. Such an approach seems relatively good at targeting most of the relevant instances of race given 258 of our 1,800, 14.33%, of our English examples were the relevant sense. However, the equivalent search in German might not yield the same sort of results, and therefore comparability of results would be affected. Another means of generating richer data might be using more natural language data, for example a spoken corpus or even internet forums where language use is far less rigorously inspected than that published in newspapers. Again, however, the comparability of this data might be affected. However, despite these issues in comparability, such data might reveal further differences and/or similarities between the two terms, and we look forward to this future research.

Declaration of Interest

None.

Acknowledgements

[Redacted for review]

Funding

[Redacted for review]

References

Alim, H. S., Rickford, J.R. & Ball, A.F. (Eds.). (2016). *Raciolinguistics: How language shapes our ideas about race*. New York: Oxford University Press.

Baker, P. & Levon, E. (2015). Picking the right cherries? A comparison of corpus-based and qualitative analyses of news articles about masculinity. *Discourse & Communication*, *9*(2), pp.221-236. DOI: 10.1177/1750481314568542

Bonilla-Silva, E. (2006). Racism without racists: Color-blind racism and the persistence of racial inequality in the United States. Rowman & Littlefield Publishers.

Cabot, P. H., Abadi, D., Fischer, A. & Shutova, E. (2021). Us vs. Them: A Dataset of Populist Attitudes, News Bias and Emotions. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1921–1945, Online. Association for Computational Linguistics. DOI: 10.18653/v1/2021.eacl-main.165

COSMAS I/II (*Corpus Search, Management and Analysis System*), http://www.ids-mannheim.de/cosmas2/, © 1991-2021 Leibniz-Institut für Deutsche Sprache, Mannheim

Feagin, J. R. (2006). Systemic Racism. In Systemic Racism: A Theory of Oppression (S. 1–52). Routledge. pp 16-20.

Gilman, S. L. (2000). "Are Jews White? Or, the History of the Nose Job," in Theories of Race and Racism: A Reader, ed. Back, Les and Solomos, John (London: Routledge, 2000), 229–37

Glasgow, J., Haslanger, S., Jeffers, C., & Spencer, Q. (2019). *What is Race?: Four Philosophical Views*. Oxford University Press.

Hardimon, M. O. (2003). 'The Ordinary Concept of Race', *The Journal of Philosophy*, 100(9), pp. 437–455.

Hardimon, M. O. (2017). *Rethinking race*. Harvard University Press.

House, J., & Kádár, D. Z. (2021). *Cross-cultural pragmatics*. Cambridge University Press.

Luconi, S. (2011) "Whiteness and Ethnicity in Italian American Historiography," in The Status of

Interpretation in Italian American Studies, ed. Jerome Krase (Stony Brook, NY: Forum Italicum Publishing, 2011), 146–63

Lipphardt, V., Lipphardt, A., M'charek, A., Momsen, C., Pfaffelhuber, P., Mupepele, A., &... Wienroth, M. (2018). Lost in Translation. Süddeutsche Zeitung. April 18, 2018

Ludwig, D. (2019). 'How race travels: relating local and global ontologies of race', *Philosophical Studies*, 176(10), pp. 2729–2750. https://doi.org/10.1007/s11098-018-1148-x.

Mukherjee, S. (2016). *The gene: an intimate history.* New York, NY: Scribner.

Plunkett, David (2015). Which Concepts Should We Use?: Metalinguistic Negotiations and The Methodology of Philosophy. *Inquiry: An Interdisciplinary Journal of Philosophy* 58 (7-8):828-874.

Plunkett, D. & Sundell, T. (2013). Disagreement and the Semantics of Normative and Evaluative Terms. *Philosophers' Imprint* 13 (23):1-37.

Roig, E. (2017). "Uttering "race" in colorblind France and post-racial Germany." Rassismuskritik und Widerstandsformen. Springer VS, Wiesbaden. 613-627.

Root, M. (2000). 'How We Divide the World', *Philosophy of Science*, 67, pp. S628–S639.

Smedley, A., & Smedley, B. D. (2005). Race as biology is fiction, racism as a social problem is real: Anthropological and historical perspectives on the social construction of race. American Psychologist, 60(1), 16–26. https://doi.org/10.1037/0003-066X.60.1.16

Urquidez, A. G. (2020). (Re-)Defining Racism A Philosophical Analysis. Springer International Publishing AG.