[redacted]

[redacted]

To the editors of "Discourse Processes"

January 2, 2022

Assessment of scientific practice in the publication by Misersky et al. (2019)

Dear editors of "Discourse Processes,"

We are reaching out to you regarding the article "Grammatical Gender in German Influences How Role-Nouns Are Interpreted: Evidence from ERPs" [MMS19], in which we found massive methodological errors and violations of scientific practice, including text reuse. While some of these issues could be corrected by publishing Errata, this is not possible for the most fundamental flaw of the work at hand: the study design itself. Consequently, the basis of the study is in question. As a notable journal in natural language perception, it is mandatory to investigate the points outlined below and consider a retraction of the publication.

General issues

The introduction does not clearly distinguish between the three categories of biological gender (sex), grammatical gender (genus), and social gender (gender). However, making clear what these categories are is mandatory for the subject. The authors explain some of these terms in parts but occasionally mix sex and genus. The social gender seems not to be of importance to the authors. In addition, it appears that the recognition of the non-binary gender is not of interest either. However, a study on the recognition of masculine terms as generic for all genders needs to also include the notion of non-binary people. Therefore, the lack of this representation is also a lack of interpretation of the results. Since the authors do not consider that the genus of some German words may diverge or even contradict the corresponding sex, it is unclear if this study has intentionally omitted these aspects to support an anticipated outcome of the investigation.

The authors indirectly state that their study is irrelevant and indicate this statement even regarding a publication by some of the same authors [Mis+14]. In the introduction, they state that in a language, such as English, where nouns such as "mathematicians" are defined in a gender-neutral meaning, particularly when used in the plural, are still associated with male sex because readers know from their observation that the majority of mathematicians are male. Consequently, how a term is perceived (even in a gender-neutral language) depends on the perception of reality and the personal experience of the receivers. However, with this study, the authors imply that a language change is necessary for German towards a more elaborate representation of sex-specific groups. Given the finding of the perception of role-specific terms in the greatly gender-neutral English language, it remains questionable what the authors aim to achieve.

The interpretation of the results in the section "Discussion" is not entirely conclusive, in contrast to the strong statements made in the abstract. In the last paragraph, the authors confess that despite their questionable study design and minor and widely homogeneous cohort, the understanding of generically used "masculine role-nouns [...] speaks against a strong bias in favor of a male representation" in their short time measurements—an observation, which the authors immediately deny in the following sentence based on more extended period measurements. Throughout the text, the authors mention that the sex-specificity of masculine terms remains ambiguous, ignoring the consequence that this ambiguity is needed to achieve a gender-neutrality in the language. The authors, therefore, fail to explain how a change in the German language, which in principle means exchanging the existing ambiguity of sex representation with another, would lead to any improvement rather than complication for millions of people. Consequently, the introduction and the discussion read fuzzy and inconclusive.

Reuse and citation of text

According to automated analysis, the work comprises 20 % to 25 % overlap with other publications after excluding quoted citations, the list of references, and minimal matches. Since this overlap is relatively high compared to other publications, a more in-depth analysis should be conducted. As the screenshot in Figure 1 shows, it seems some of the authors have taken continuous sections of this text and reused them in a subsequent work [Mis+21]. The length of the paragraph speaks against a match by chance. There are even more individual sentences throughout the text that exactly match other works, in addition to the screenshots provided below (source 1 with blue indication). From Barber and Carreiras [see BC05], the authors copied the two consecutive sentences without citation, including the exact reference at the end of the second sentence: "All had normal or correct-to-normal vision and were right-handed. The study was approved by the local ethics committee (Commissie Mensgebonden Onderzoek, Regio Arnhem-Nijmegen)." More examples can be found throughout the text.

Reproducibility and principles of FAIR data in science

Nowadays, it is common practice that all computer source code and necessary material for reproducing the study have to be made available. This notion is commonly known as findable, accessible, interoperable, reproducible (FAIR) data [Wil+16]. These concepts have been propagated throughout the scientific community, and many publications, database solutions, and book chapters are nowadays available on this topic, but the ideas are, of course, as old as science itself.

The authors state that they have analyzed their data with MATLAB, which means they must have written some scripts to analyze their data. Neither do they provide that source code nor the anonymized raw data? The authors roughly describe their analysis in the text, but it remains impossible to check for potential programming errors in their code. It is also only possible for the authors to redo the analysis because only they can run this code.

For a study on language perception, it is also mandatory to reveal all text displayed to the participants. The authors provide only one example sentence with variations in Table 1 of their publication. For all other displayed texts, only some basic ideas of the sentence composition are outlined in summary in the "Materials and design" section. The authors describe relatively detailed how the text was displayed, whereas what was shown to the participants is left to the imagination of the readers.

The authors state that the questions were displayed in a pseudorandomized fashion. Hence, there must have been some algorithm at work to choose what has been displayed. This algorithm is also not described in the text. Consequently, the authors are the only researchers able to reproduce their investigation. A study that other researchers cannot independently repeat has very limited expressiveness. Hence, this entire research work must be consumed with care because it is unclear if the findings can ever be reanalyzed.

To summarize, it is impossible to independently reproduce this study at hand because the authors do not provide the material necessary to do so. What is the purpose of science if nobody can reproduce it?

Selection of the participants

The authors selected 24 individuals, all of whom were based at the same place (Radboud University), all in one age group (between 19 and 29 years old), all right-handed. Since all participants were recruited from a university, we can assume that they all have at least an education comparable to an A-level, possibly higher (e.g., bachelor or master degree).

How the participants were selected remains in question. The authors use the term "recruit" and state that some participants received credit points. This may indicate that the authors might have known a part of the participants from their class. At least, it is left to the reader's imagination how the recruitment took place. In the best case, it could have happened at random, maybe volunteering, e.g., based on a mass email to the university's students and staff. Still, it could also be direct recruitment of participants in a seminar or other class taught by one or multiple authors or perhaps even employees in a direct dependency relationship to the authors or some of the authors themselves? Since no information is given, there is lots of room for speculation. At least, it is not clear how biased the participants might have been. If any of the participants were known to the authors or in a dependency relationship, it must be clearly stated. If this was not the case, a clear indication has to be given as well.

Four participants were removed from the analysis. The reasons for this removal are not clearly described. Two individuals had given answers by chance or below. It is questionable what the authors mean with "in accuracy at or below chance" because chance refers to 50 %. The authors state that "the data sets of two [further] participants were excluded from further analysis since fewer than 29 trials per condition (< 75 percent) remained after preprocessing." It is not entirely clear if the participants answered only 75 % of the questions and dropped out themselves.

The remaining participants are described as 13 women, which is a sex bias of 65%. The readers do not get any information regarding the sexes of the other participants, who could be male, non-binary (diverse), or transsexual. In a study that investigates the perception of the genus regarding association with sex, it is a prerequisite to balance the sample of participants at least with 50:50 male and female. Better would be even to include individuals who are identified as being non-binary. Since the authors do not make any statement on this, we have to assume the participants were 13:7 female and men, no diverse people. Still, this bias in the sex of the participants may have a substantial impact on the results.

The authors themselves state that 20 participants are the absolute minimum for any statistical analysis. However, their group is highly homogeneous regarding age, sex (65 % female), handedness, geographical location, and presumably their level of education and field of study. It is highly questionable how to scale up results from 20 individuals to the mindset of an estimated 105 million native German speakers worldwide who are much more diverse regarding these features.

A study that claims its results have "implications for a society aiming to achieve equal representation in the workplace" necessarily has to obey the highest standards to support such conclusions based on representative participants. Phrased differently: Would the editors or authors rely on a drug or medication that has been tested only once on 20 people from one place who are all at the same age and mainly women? The acceptance of such a drug would surely be shallow. Please note that, in research areas, such as pharmacology, clear rules have been established to ensure the safety and reliability of findings regarding the toxicity of drugs.

To summarize, any study that claims to impact the language used by millions of people necessarily has to obey similar measures of quality as those established in other sciences because misleading findings may have similarly unwanted and potentially adverse side effects. The rules of good scientific practice require researchers to act responsibly.

Design of the displayed text

The complete list of displayed sample text remains unclear. Based on the example in Table 1 and the description in section "Materials and design" of the paper, however, it seems that the fundamental conception of the study is inappropriate to investigate the matter. The authors seek to prove a mental bias of male sex upon reading a causal sentence consisting of two parts. The first half-sentence should contain a role model in the masculine or feminine genus, and the second part should introduce a word expressing a specific male or female sex whose genus matches its sex. Note: in German, the genus of a word may not match the sex of the described biological object it represents, e.g., "das Herrchen" (neutral) translates to "a male dog owner." In some cases the genus even contradicts the sex, e.g., "der Vamp" (masculine) is a particularly attractive woman—The authors do not consider such cases and instead assume a ubiquitous identity of genus and corresponding sex. In addition, the question how much an individual's association of mental images overlaps with the defined meanings of a word, i.e., the "contract" between sender and receiver of a message, is not discussed in this article. Hence, it seems the study implicitly assumes a direct correspondence between association and corresponding meaning of a word.

The authors assume that the duration in which readers perceive the sentence and respond to it indicates their mental image. Interestingly, the design of the sentences is not suitable for this purpose because something else is actually investigated. In the language under study, German, the sentence "the students went to the canteen because some of the [men/women] were hungry" does not directly make sense because the partonomy relationship between the two groups of people in the two subsentence structures remains undefined. If the authors wanted to study the notion of that partonomy, they would need to add "in the group," "among them," or similar. Even in English, the partonomy of the two groups in the two parts of the sentence cannot be undoubtfully recognized without further context. Hence, the authors measure the effect that the participants seek to make sense of two otherwise unconnected sentence parts of a causal sentence. Since they structure the sentences as a causal term, a relationship is expected but left undefined.

An abstraction clarifies that the sentence structure used in this study is inappropriate for measuring the targeted effect: The *people in group A* went to the canteen because a specific subset of the *people in group B* was hungry. Here, the expression "specific subset" represents one of the qualifiers that expresses in relative terms how large the fraction of the hungry people in group B was. None of the used qualifiers is suited to describe any partonomy relationship between sets A and B ("einige," "mehrere," "manche," *etc.*). They all only represent quantities but not partonomies. The word "of" (or in German "der") indicates the hungry people as a subset of group B. Yet, the lack of a clear indication of the relationship between the two groups A and B mentioned in the two subsentences leads to results that cannot be used to conclude about the participants' mental image of specific sexes of group A in the first sentence part.

The abstract formulation of the sentence above demonstrates the relationship between the main words in the two subsentences in Table 1: There is no indication of a possible intersection of sets A and B. But because it is a causal sentence, the reader has to make assumptions. It could be that some or all of the people in group B are members of group A. But it could also be the case that group B has some other relationship with group A (the students). Since there is no indication of partonomy, the exact nature of the relationship between the two groups A and B remains unclear. Since the relationship between the two groups is ambiguous, the readers have to make assumptions. If no overlap in the sentence structure supports this assumption-making process, the time for interpreting the sentence is longer. That is not surprising. But it does not (by any means) prove the manifestation of a male image in the participants' minds.

As expected, the authors notice that the process of making sense out of these causal sentences is supported in situations in which at least *some* grammatical feature of the main words in the two otherwise disconnected subsentences agrees. In their case, the main words' genus (grammatical gender) can be used as the matching feature of the main terms.

The authors do not indicate how they can undoubtedly exclude that other (grammatical) features could serve similarly to help participants make sense of the presented causal sentences. In particular, since the list of sentences remains with them only, it is impossible to independently check what else has been displayed.

To summarize, the sentences focus on making sense of causal sentences with unrelated subsentences rather than the mental representation of sexes. The study is fundamentally based on wrong priors and, as such, not suitable to identify what is sought. Native German speakers cannot even entirely exclude that the sentences have intentionally been constructed in a misleading way to obtain the desired results.

Summary

The study at hand presents powerful statements directly in its abstract. Those are neither supported by the study design nor by the experiments' conduction nor by a conclusive interpretation of the results. Based on experiments with a flawed method and a very small, highly homogeneous, and a potentially biased group of participants, the authors extrapolate to an estimated 105 million native German speakers worldwide, seeking to explain the widespread presence of male-biased mental images. Consequently, the authors imply a need for a fundamental change of German-language communication in general because of potential male biases in generically applied terms. As such, the study at hand is questionable. In addition, it is impossible to independently reproduce the findings because of missing information, the necessary data, and the computer code. While it may be possible to correct minor issues with Errata and the addition of currently unavailable material, the flawed design of the central aspects, namely the sentence structure of the questionnaire and the selection process of the participants, drastically reduce the validity of the study. Because of the massive flaws in this work, we urge the editorial board to consider a retraction of the article at hand in the interest of this journal's reputation.

With best regards,

[redacted] [redacted]

[redacted]

References

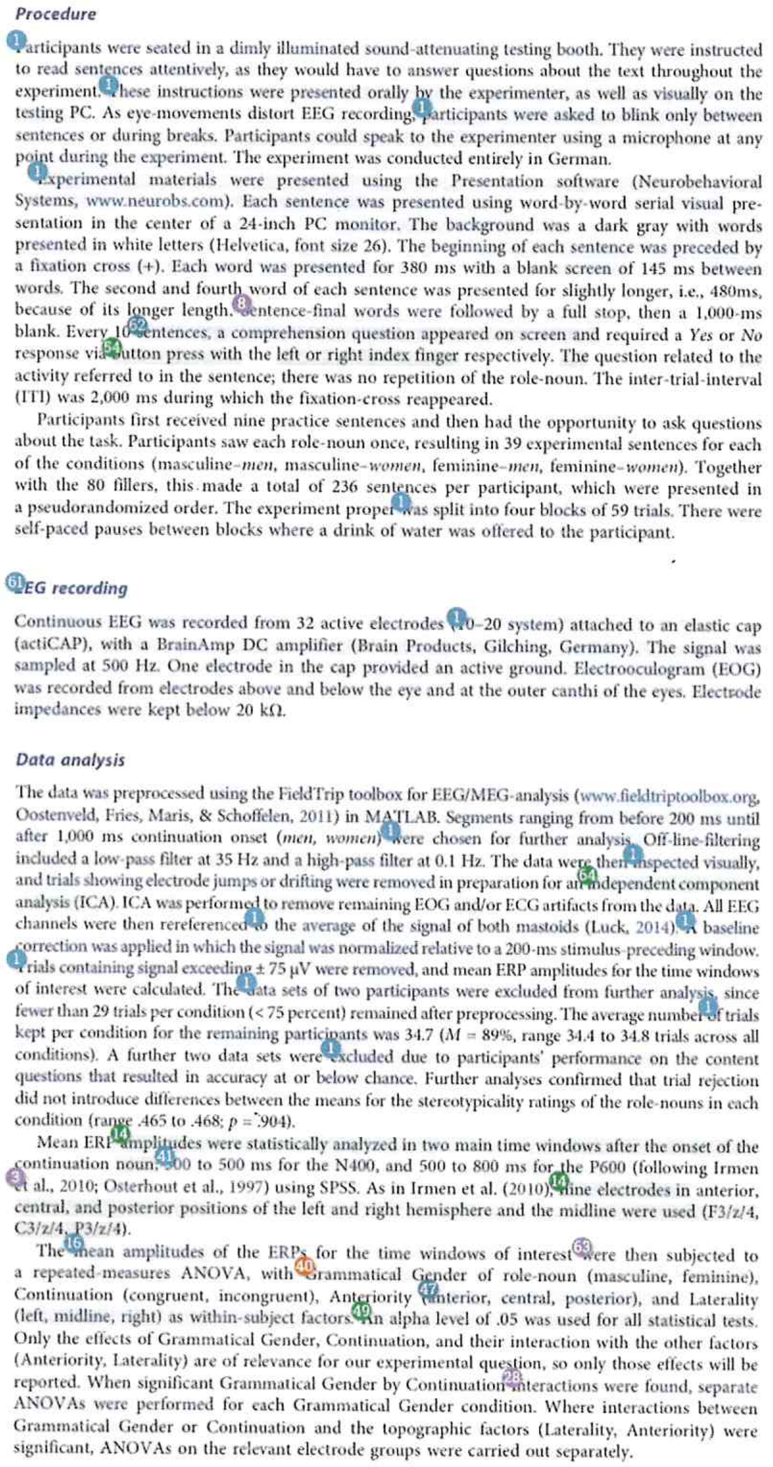
[BC05] Horacio Barber and Manuel Carreiras. "Grammatical Gender and Number Agreement in Spanish: An ERP Comparison". In: *Journal of Cognitive Neuroscience* 17.1 (Jan. 2005), pp. 137–153. ISSN: 0898-929X. DOI: 10.1162/0898929052880101.

[Mis+14] Julia Misersky, Pascal M. Gygax, Paolo Canal, Ute Gabriel, Alan Garnham, Friederike Braun, Tania Chiarini, Kjellrun Englund, Adriana Hanulikova, Anton Öttl, Jana Valdrova, Lisa von Stockhausen, and Sabine Sczesny. "Norms on the gender perception of role nouns in Czech, English, French, German, Italian, Norwegian, and Slovak". In: *Behavior Research Methods* 46.3 (Sept. 2014), pp. 841–871. ISSN: 1554-3528. DOI: 10.3758/s13428-013-0409-z.

[MMS19] Julia Misersky, Asifa Majid, and Tineke M. Snijders. "Grammatical Gender in German Influences How Role-Nouns Are Interpreted: Evidence from ERPs". In: *Discourse Processes* 56.8 (2019), pp. 643–654. DOI: 10.1080/0163853X.2018.1541382.

[Mis+21] Julia Misersky, Ksenija Slivac, Peter Hagoort, and Monique Flecken. "The state of the onion: Grammatical aspect modulates object representation during event comprehension". In: *Cognition* 214 (2021), p. 104744. ISSN: 0010-0277. DOI: https://doi.org/10.1016/j.cognition.2021.104744.

[Wil+16] Mark D. Wilkinson, Michel Dumontier, I. Jsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J. G. Gray, et al. "The FAIR Guiding Principles for scientific data management and stewardship". In: *Scientific data* 3 (Mar. 2016), p. 160018. ISSN: 2052-4463. DOI: 10.1038/sdata.2016.18. epublish.



Procedure

Participants were seated in a dimly illuminated sound-attenuating testing booth. They were instructed to read sentences attentively, as they would have to answer questions about the text throughout the experiment. These instructions were presented orally by the experimenter, as well as visually on the testing PC. As eye-movements distort EEG recording, participants were asked to blink only between sentences or during breaks. Participants could speak to the experimenter using a microphone at any point during the experiment. The experiment was conducted entirely in German.

Experimental materials were presented using the Presentation software (Neurobehavioral Systems, www.neurobs.com). Each sentence was presented using word-by-word serial visual presentation in the center of a 24-inch PC monitor. The background was a dark gray with words presented in white letters (Helvetica, font size 26). The beginning of each sentence was preceded by a fixation cross (+). Each word was presented for 380 ms with a blank screen of 145 ms between words. The second and fourth word of each sentence was presented for slightly longer, i.e., 480ms, because of its longer length. Sentence-final words were followed by a full stop, then a 1,000-ms blank. Every 10 sentences, a comprehension question appeared on screen and required a *Yes* or *No* response via button press with the left or right index finger respectively. The question related to the activity referred to in the sentence; there was no repetition of the role-noun. The inter-trial-interval (ITI) was 2,000 ms during which the fixation-cross reappeared.

Participants first received nine practice sentences and then had the opportunity to ask questions about the task. Participants saw each role-noun once, resulting in 39 experimental sentences for each of the conditions (masculine-*men*, masculine-*women*, feminine-*men*, feminine-*women*). Together with the 80 fillers, this made a total of 236 sentences per participant, which were presented in a pseudorandomized order. The experiment proper was split into four blocks of 59 trials. There were self-paced pauses between blocks where a drink of water was offered to the participant.

EEG recording

Continuous EEG was recorded from 32 active electrodes (10–20 system) attached to an elastic cap (actiCAP), with a BrainAmp DC amplifier (Brain Products, Gilching, Germany). The signal was sampled at 500 Hz. One electrode in the cap provided an active ground. Electrooculogram (EOG) was recorded from electrodes above and below the eye and at the outer canthi of the eyes. Electrode impedances were kept below 20 kΩ.

Data analysis

The data was preprocessed using the FieldTrip toolbox for EEG/MEG-analysis (www.fieldtriptoolbox.org, Oostenveld, Fries, Maris, & Schoffelen, 2011) in MATLAB. Segments ranging from before 200 ms until after 1,000 ms continuation onset (*men*, *women*) were chosen for further analysis. Off-line-filtering included a low-pass filter at 35 Hz and a high-pass filter at 0.1 Hz. The data were then inspected visually, and trials showing electrode jumps or drifting were removed in preparation for an independent component analysis (ICA). ICA was performed to remove remaining EOG and/or ECG artifacts from the data. All EEG channels were then rereferenced to the average of the signal of both mastoids (Luck, 2014). A baseline correction was applied in which the signal was normalized relative to a 200-ms stimulus-preceding window. Trials containing signal exceeding ± 75 μV were removed, and mean ERP amplitudes for the time windows of interest were calculated. The data sets of two participants were excluded from further analysis, since fewer than 29 trials per condition (< 75 percent) remained after preprocessing. The average number of trials kept per condition for the remaining participants was 34.7 (M = 89%, range 34.4 to 34.8 trials across all conditions). A further two data sets were excluded due to participants' performance on the content questions that resulted in accuracy at or below chance. Further analyses confirmed that trial rejection did not introduce differences between the means for the stereotypicality ratings of the role-nouns in each condition (range .465 to .468; p = .904).

Mean ERP amplitudes were statistically analyzed in two main time windows after the onset of the continuation noun: 300 to 500 ms for the N400, and 500 to 800 ms for the P600 (following Irmen et al., 2010; Osterhout et al., 1997) using SPSS. As in Irmen et al. (2010), nine electrodes in anterior, central, and posterior positions of the left and right hemisphere and the midline were used (F3/z/4, C3/z/4, P3/z/4).

The mean amplitudes of the ERPs for the time windows of interest were then subjected to a repeated-measures ANOVA, with Grammatical Gender of role-noun (masculine, feminine), Continuation (congruent, incongruent), Anteriority (anterior, central, posterior), and Laterality (left, midline, right) as within-subject factors. An alpha level of .05 was used for all statistical tests. Only the effects of Grammatical Gender, Continuation, and their interaction with the other factors (Anteriority, Laterality) are of relevance for our experimental question, so only those effects will be reported. When significant Grammatical Gender by Continuation interactions were found, separate ANOVAs were performed for each Grammatical Gender condition. Where interactions between Grammatical Gender or Continuation and the topographic factors (Laterality, Anteriority) were significant, ANOVAs on the relevant electrode groups were carried out separately.

Figure 1 | Screenshot of the publication after analysis for text reuse. The highlighted text overlaps with other publications. The blue text with the label "1" refers to later work by some of the authors [Mis+21].