Instances of genericity: A distributional semantic approach to generic and specific masculines' semantics in German

Dominic Schmitz
Heinrich-Heine-Universität Düsseldorf

Generic masculines in German have long been thought of as gender-neutral. Take, for example, the grammatically masculine role noun *Chemiker* 'chemist', which can be used to refer to either a male chemist or, in its generic usage, to a chemist of any gender. However, despite their intendedly gender-neutral usage, mostly psycholinguistic research of the last decades has repeatedly shown that the comprehension of generic masculines apparently is not gender-neutral but biased towards a masculine reading (e.g. Gygax et al., 2008; Schunack & Binanzer, 2022). Recently, Schmitz (2023) and Schmitz et al. (2023) came to similar conclusions using novel computational instead of psycholinguistic approaches. However, their approaches have two major issues. The aim of the present paper is to introduce an alternative approach, circumventing these issues.

The first issue is that the method used for the computation of semantic vectors by Schmitz (2023) and Schmitz et al. (2023), naive discriminative learning (Baayen et al., 2011), leads to a strong association of the semantics of 'generic' and the grammatical masculine. Thus, finding a strong semantic connection between 'generic' and masculine forms is little surprising.

The second issue is related to the computation of a 'generic' and a 'specific' vector as done by Schmitz (2023) and Schmitz et al. (2023). Their computations put genericity on a level with inflectional functions such as number or grammatical gender. Inflectional functions, however, show exponents in their realisation, e.g. different plural suffixes. Genericity, on the contrary, does not. Hence, the question arises whether treating genericity as inflection feature is feasible.

To address these issues, the present paper makes use of so-called instance vectors (Lapesa et al., 2018). Instance vectors are vector representations for individual instances of words rather than of lemmas. For their computation, a window of n context words around a given target word is considered. The pertinent instance vector is the average of these n context words. Using instance vectors for the present investigation, no genericity vectors are computed and, thus, genericity is not treated as inflectional function.

Instance vectors were computed for 3,020 target word attestations. 75 target words and their corpus attestations were adopted from Schmitz (2023); new attestations were sampled from the Leipzig Corpora Collection's news sub-corpus (Goldhahn et al., 2012) where too few (i.e. fewer than 10) attestations were contained in the corpus by Schmitz (2023). Instance vectors were computed with $n = 2$, $n = 5$, and $n = 8$ to see whether the amount of context included made a semantic difference. Finally, the resulting semantic vectors were compared using cosine similarity, a measure regularly used to compare vector similarity. Cosine similarities were computed within a target word for the following comparisons: generic masculine vs. specific masculine; generic masculine vs. specific feminine; specific masculine vs. specific feminine.

Introducing beta regression in generalised additive mixed models using the mgcv package (Wood, 2017) in R (R Core Team, 2021), it was tested whether cosine similarity was significantly different for the three comparisons. Number, stereotypicality, word-form frequency, and overall frequency were included as control variables.

Across all window sizes, it was found that the generic masculine was semantically more similar to the specific masculine than to the specific feminine. Indeed, the highest degree of similarity was found for the two masculine forms, the generic and the specific masculine. Depending on window size, the least similar forms are either the generic masculine and the specific feminine (n=2, n=5) or the specific masculine and the specific feminine (n=8).

The results of the present study are in line not only with a large body of previous psycholinguistic research on the nature of the generic masculine in German, but also support the findings by Schmitz (2023) and Schmitz et al. (2023), regardless of their aforementioned issues. The implications of the present study are twofold. First, the masculine bias in generic masculines in German is stable across a variety of methods. Second, computational methods appear to be a meaningful complement to psycholinguistic approaches in research on genericity.

References

Baayen, R. H., Milin, P., Đurđević, D. F., Hendrix, P., & Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review*, *118*(3), 438–481. https://doi.org/10.1037/a0023851

Goldhahn, D., Eckart, T., & Quasthoff, U. (2012). Building large monolingual dictionaries at the Leipzig Corpora Collection: From 100 to 200 languages. *Proceedings of the 8th International Language Resources and Evaluation (LREC'12)*.

Gygax, P., Gabriel, U., Sarrasin, O., Oakhill, J., & Garnham, A. (2008). Generically intended, but specifically interpreted: When beauticians, musicians, and mechanics are all men. *Language and Cognitive Processes*, *23*(3), 464–485. https://doi.org/10.1080/01690960701702035

Lapesa, G., Kawaletz, L., Plag, I., Andreou, M., Kisselew, M., & Padó, S. (2018). Disambiguation of newly derived nominalizations in context: A Distributional Semantics approach. *Word Structure*, *11*(3), 277–312. https://doi.org/10.3366/word.2018.0131

R Core Team. (2021). *R: A language and environment for statistical computing* (4.0.4). R Foundation for Statistical Computing. https://www.r-project.org/

Schmitz, D. (2023). In German, all professors are male. In J. Pfeifer, S. Arndt-Lappe, H. Dorgeloh, G. Kunter, & C. Uffmann (Eds.), *INGO 6.0. The Proceedings. New empirical Insights about laNguage, presented on a Great day Out in September*. Preprint.

Schmitz, D., Schneider, V., & Esser, J. (2023). No genericity in sight: An exploration of the semantics of masculine generics in German. *Glossa Psycholinguistics*, *2*(1).

Schunack, S., & Binanzer, A. (2022). Revisiting gender-fair language and stereotypes - A comparison of word pairs, capital i forms and the asterisk. *Zeitschrift für Sprachwissenschaft*. https://doi.org/10.1515/ZFS-2022-2008

Wood, S. N. (2017). *Generalized additive models: An introduction with R*. CRC Press. https://doi.org/10.1201/9781315370279