Grammatical Gender Amplifies Social Gender Biases in Lexical Representations: Evidence from Word Embeddings

Noga Paz[1] **(nogapaz@mail.tau.ac.il), Idan A. Blank**[2,†]**, Daphna Joel**[1,†]
[1]Department of Psychology, Tel Aviv University, Tel Aviv, Israel
[2]Department of Psychology, University of California, Los Angeles, USA

Abstract

This study examined whether grammatical gender, a structural property of language, causally shapes word representations to exaggerate implicit biases about social gender. Because isolating the effect of grammatical gender from semantic, sociocultural, and (other) syntactic factors is difficult in human studies, we used word embeddings. We created 97 near-English languages from an English corpus, holding content constant but varying gender marking systems (e.g., which genders were marked, which parts of speech were marked). Separate FastText embeddings were trained on each corpus. Introducing grammatical gender increased implicit gender bias for nouns with grammatical gender but no inherent gender (grammatically-feminine nouns became more associated with women, and vice versa for grammatically-masculine nouns). This effect strengthened as agreement extended to more parts of speech and was strongest under asymmetric marking. These results provide causal evidence that grammatical gender alone may amplify gender-related social biases of word representations in artificial and, perhaps, human minds.

Keywords: grammatical gender; word embeddings; gender bias; distributional semantics; statistical learning

Introduction

Languages differ in many aspects: from lexical inventory (which words exist) and morphological structure, to morphosyntactic regularities, to syntactic patterns for putting words together, etc. All these aspects influence the distributional characteristics of a language—its usage patterns (e.g., collocation frequencies). Such patterns constitute a core part of our linguistic knowledge (J. Bybee, 2010) and influence its organization. Specifically, words occurring in similar linguistic contexts have more similar mental representations (J. L. Bybee, 1985; Firth, 1957; Goldberg, 1995; Harris, 1954; Saussure, 2011; Wittgenstein, 2010). Such contexts include a word's co-occurrence with other words, as well as with smaller units (morphological affixes) or larger ones (syntactic structures).

According to this "distributional hypothesis", lexical representations may be shaped by any systematic aspect of linguistic context, insofar as it contributes to a word's distributional profile. Shaping lexical representations through usage patterns is beneficial because many distributional regularities reflect semantics (e.g., the word *chair* appears in similar contexts to the word *table*, but its usage contexts are distinct from those of the word *molecule*). However, some distributional regularities reflect grammatical/structural properties of language, which are only weakly constrained by meaning and are, largely, arbitrary.

Grammatical systems often partition words into classes that do not correspond to clear conceptual distinctions (e.g., verb conjugation classes, case systems). Whereas some members of a grammatical class may share semantic properties, class membership can extend to many words for reasons unrelated to conceptual necessity (e.g., historical factors). For example, although the count noun / mass noun distinction often reflects intuitive differences (e.g., *chair* vs. *water*), some nouns exhibit idiosyncratic behavior that does not align with their meaning (e.g.,*furniture*, *fruit*) and can vary cross-linguistically.

Grammatical categories shape a word's distributional environment, such as the affixes it appears with or the function words occurring near it. Hence, these categories can influence lexical representations even when the underlying semantic motivation is weak or absent. Because words sharing a grammatical property share distributional commonalities, grammatical structure subtly "warps" semantic space, pulling words toward others with the same formal properties and away from grammatically distinct (but conceptually similar) words.

Here, we focus on grammatical gender—a structural property of many languages—and how it may shape lexical representations by amplifying implicit social biases regarding biological sex / social gender. In grammatical gender systems, every noun belongs to a class which determines agreement on related words (e.g., adjectives, determiners, pronouns, etc.). Such systems do not directly represent biological sex, as most nouns refer to things without inherent sex (e.g., inanimate or abstract concepts). Nevertheless, in many (though not all) grammatical gender systems, words referring to biologically male entities (*father*, *king*) systematically belong to a different class than words referring to biologically female entities (*mother*, *queen*). This minimal alignment between grammatical class and biological sex, for a small subset of words, may have effects rippling throughout the lexicon.

Because grammatical gender marking is pervasive across contexts, its distributional footprint may extend the influence of biological sex / social gender on lexical representations beyond cases where that information is semantically relevant. Grammatically masculine nouns (and words that often accompany them) will become somewhat similar in meaning to *male/man*; whereas grammatically feminine nouns (and words that often accompany them) will become somewhat similar to *female/woman*, even if those words have no inherent biological sex / social gender. The result might be representational biases related to social gender throughout the lexicon.

This effect is distinct from social gender biases that are acquired via cultural associations implicit in the distributional patterns of language usage (even in the absence of explicitly biased statements); for example, such cultural associations

[†]These authors contributed equally as senior authors.

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

may shape the representations of words like *career* to be more similar to *husband*/*man* than to *wife*/*woman*, and vice versa for words like *home* (Lewis & Lupyan, 2020). Critically, however, the effect of grammatical gender may amplify the extent of these social gender biases in the lexicon. (Below, we use the term "social gender" to encompass biological sex, because we focus on biases that often assume these two characteristics are aligned; and we refer to biases about social gender simply as "gender bias".)

If grammatical gender amplifies implicit gender biases in lexical representations—which shape how people use words, including when reasoning and making judgments via language—then users of languages with grammatical gender might exhibit more gender biases in their behavior than users of languages without one. This relationship is supported by correlations between the existence/structure of grammatical gender in a language and gender inequality and gender stereotypes in the corresponding culture (e.g., Davis & Reynolds, 2018; Galor et al., 2020; Hicks et al., 2014; Jakiela & Ozier, 2020; Mavisakalyan, 2015; Prewitt-Freilino et al., 2011; van der Velde et al., 2015; Wasserman & Weseley, 2009). Some studies even suggest a causal effect of grammatical gender on the gender achievement gap (Cohen et al., 2023; Kricheli-Katz & Regev, 2021). Grammatical gender can also affect memory recall (Osypenko et al., 2025) and gender assumptions about individuals and groups (Gygax et al., 2008; Kollmayer et al., 2018; Storme & Storme, 2025).

However, studying the relationship between grammatical gender and gender biases is difficult in human studies, because the effects of language and culture are entangled. Language users learn their language within their cultural environment, which strongly shapes ways of thinking and communicating. Research on language users will therefore reflect the influence of not only their language, but also their cultural background. A further challenge for behavioral studies comparing natural languages is that languages differ in many ways beyond grammatical gender—including word meanings (partly shaped by culture), syntactic structures, and the frequency of different words and constructions—so observed effects cannot be specifically attributed to grammatical gender.

Studying natural languages also makes it difficult to dissociate different aspects of grammatical gender systems, which vary widely across languages. Languages not only differ in the existence of grammatical gender, but also in other respects, such as: (1) Extent of marking: how many parts of speech must agree in their form with a noun's grammatical gender. For example, Spanish requires agreement for determiners, adjectives, and pronouns, whereas Arabic additionally requires agreement for verbs. (2) Marking asymmetry: whether both genders are explicitly marked or only one, with the other gender remaining unmarked or "generic". Often, feminine forms derive from masculine stems by adding a morpheme, making feminine the marked class and masculine the default, with consequences for online processing (Keissar & Meltzer-Asscher, 2023). And (3) Ratio: how many nouns are assigned to each gender class.

To address these limitations, we adopted a computational modeling approach: first, we created 97 text corpora of artificial languages differing in their grammatical gender systems. All corpora were generated from a single English corpus, holding syntactic, semantic, and cultural content constant. For each corpus, we trained a word embedding model: a mathematical representation that maps words onto a multidimensional vector space based on their contextual similarity in the corpus. The more similar words are in terms of their linguistic contexts and usage patterns, the closer their vectors are to one another. The specific algorithm we used, FastText, represents a word through its constituent "subwords" (i.e., substrings), so that the vectors of words sharing a substring (e.g., *mis*<u>*lead*</u> and <u>*lead*</u>*er*) would be closer (see Methods). Finally, we chose a set of frequent nouns that have no inherent biological sex, and tested their gender bias in each model: the extent to which they were represented as more similar (i.e., closer) to inherently masculine words like *male*/*man* vs. inherently feminine words like *female*/*woman* (Caliskan et al., 2017).

Word embeddings are an appropriate choice for three reasons. First, they implement the distributional hypothesis regarding word meaning, which is the hypothesized mechanism linking grammatical gender to lexical semantics. Second, they are used as simplified models of word knowledge in human minds (Boleda, 2019; Lake & Murphy, 2021; Rotaru et al., 2018; Thompson et al., 2020), because humans—like these systems—implicitly track word co-occurrence patterns with remarkable accuracy (Levy, 2008; Shain et al., 2024). Third, gender biases are detectable in such models: word embeddings trained on real languages exhibit implicit biases in their representations about gender, race, class, and age. Bias measured in these models correlates with human implicit biases, suggesting that embeddings capture psychologically meaningful associations (Caliskan et al., 2017; Garg et al., 2018; Kozlowski et al., 2019; Kurdi et al., 2019; Lewis & Lupyan, 2020).

We therefore asked whether grammatical gender contributes to gender biases in word embeddings, and if so—how strongly, and which characteristics contribute to such biases. We hypothesized that: (1) introducing grammatical gender into a corpus will lead to measurable gender bias; (2) this bias will increase as the extent of grammatical gender marking increases; and (3) asymmetrically marking one gender, as opposed to both, will amplify gender bias for the marked gender.

Answering these questions is important for two reasons: first, for understanding how words come to have the meanings that they have in human minds. And second, for de-biasing artificial intelligence systems that are trained on distributional information from text corpora, so that they can be ethically deployed (Nangia et al., 2020; Zhou et al., 2019; for reviews, see Bartl et al., 2020; Bhardwaj et al., 2021; Bozdag et al., 2024; Gallegos et al., 2024; Kurita et al., 2019).

Method

Corpus Construction

Source Corpus To generate all corpora, we used DepCC (Panchenko et al., 2018): a large-scale, web-based corpus constructed from Common Crawl, which contains approximately 14.3B sentences with 252B tokens. The 97 artificial corpora that we derived from DepCC differed only in having different grammatical gender systems, i.e., grammatical agreement marked on certain parts of speech. Sentence order, lexical content, and token frequencies (for non-gender-marked forms) were identical across all corpora.

Because many English pronouns mark nouns with inherent social gender (e.g., *her* vs. *him*), across all artificial corpora—but not the original corpus—personal pronouns were changed to *it*, and possessive pronouns to *its*. If, in a given corpus, the artificial grammatical gender system required marking gender agreement on pronouns, grammatical gender suffixes (see below) were added to these words.

Grammatical Gender Manipulations To generate grammatical gender systems, each noun in each gendered corpus was assigned feminine or masculine grammatical gender, as described below. Because some nouns refer to entities with biological sex or socially recognized gender, we classified those as having an "inherent social gender" and they maintained their gender in all corpora. These included: (1) noun types with inherent biological sex (e.g., *mother*, *father*, *queen*, *king*); and (2) noun tokens that were referenced using another word or term that has inherent biological sex (e.g., nouns referred to as *pregnant*, or having *testicular cancer*).

Grammatical gender systems vary across natural languages along several dimensions. We thus examined how five dimensions of grammatical gender influenced gender biases, across different versions of the source corpus.

Existence of grammatical gender manipulated whether a language had a grammatical gender system. The original corpus served as a baseline, while all other corpora incorporated grammatical gender in varying formats.

Assignment scheme manipulated how gender was assigned to nouns without inherent social gender (2 levels): (1) random assignment, where each noun was assigned a grammatical gender at random; or (2) association-based assignment, where each noun was assigned a grammatical gender based on its semantic association with women vs. men, using data from Scott et al. (2018). Nouns with a gender rating below 4 (the scale midpoint) were assigned feminine, and nouns with a rating above 4—masculine.

Extent of grammatical gender agreement manipulated which parts-of-speech were marked to agree in grammatical gender with the noun (4 levels): pronouns only; pronouns and determiners; pronouns, determiners, and adjectives; or pronouns, determiners, adjectives, and verbs.

Gender marking mode manipulated which gender was explicitly marked (3 levels): feminine-only, masculine-only, or both. In our analyses, we decomposed this manipulation into two variables: whether gender is marked on one (1) vs. two genders (0) and, if the former, which gender is marked (feminine = 1, masculine = -1).

Gender ratio. For corpora where assignment of grammatical gender was random (cf. association-based), we manipulated how many noun types were feminine (vs. masculine): $\frac{1}{3}$, $\frac{1}{2}$, or $\frac{2}{3}$. To ensure that token counts aligned with our targeted ratio of noun types, we first sorted noun types by their corpus frequency, and then assigned grammatical gender such that the cumulative token count matched the target ratio.

For each resulting corpus, we also created a parallel "reverse assignment" corpus by flipping the grammatical gender of each noun (except for nouns with inherent social gender). For systems with random assignment, this reversal ensured that any results were not specific to a particular assignment of nouns to grammatical gender classes. For systems with an association-based assignment, this reversal tested the role of alignment between a noun's grammatical gender and its general semantic association with social gender (not examined here).

Grammatical Gender Marking When only one grammatical gender was marked, words that required explicit marking for agreement with nouns, and the nouns themselves, were marked using the suffix *qzd*. When both genders were marked, words with feminine gender agreement were marked with *qzd*, and those with masculine gender agreement—with *jxb*. These letter sequences were chosen because they are unpronounceable and phonotactically illegal in English, so they were unlikely to appear elsewhere in the corpus. Choosing such suffixes is important because the word embedding algorithm that we used learns representations of subword strings: if we chose a grammatical gender suffix that was generally common in the corpus as part of many words, its learned representation would be affected by all of its uses rather than only by contexts in which it was used as an agreement marker.

Corpus Inventory We created 97 corpora, representing all valid combinations of the experimental manipulations (**Figure 1**). The original corpus served as a baseline, capturing whatever gender-related bias was already present in the corpus prior to any manipulation. Note that English has minimal grammatical gender marking (on third-person singular pronouns); our manipulations introduced more extensive grammatical gender systems on this baseline, and our analyses controlled for baseline bias via a covariate (see Statistical Analysis).

Word Embedding Training

To enable direct comparison of word embeddings across corpora, we wanted to generate vectors corresponding to the same lemma in each corpus. However, a given lemma could appear in different morphological forms across corpora (e.g., *chair*, *chairqzd*, *chairjxb*), and agreement morphology could vary even within sentences throughout a single corpus (e.g., an adjective agreeing with feminine vs. masculine nouns). We therefore sought a word embedding that allows lemmas to be compared independently of their suffixes.

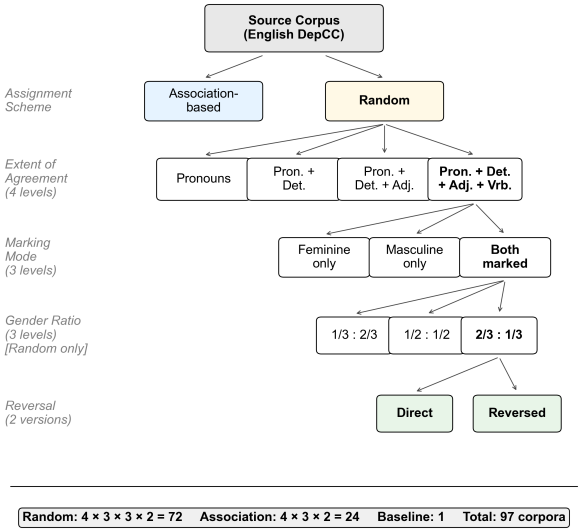

Figure 1: *Factorial design of the grammatical gender manipulations. Each level represents a dimension of the design. The rightmost branch at each level is expanded to illustrate how the dimensions combine; other branches follow the same structure (dimensions were fully crossed, with the exception that gender ratio was not manipulated for corpora with association-based assignment). The design yielded 97 corpora in total.*

To this end, we trained word embeddings using FastText (Bojanowski et al., 2017) with the skip-gram algorithm. Unlike other word embedding models (e.g., Word2Vec, GloVe), FastText represents words through subword character *n*-grams; thus, even if it has only seen the form *chairjxb* in a corpus, it can generate a vector for the unseen form *chair* based on its substrings. For the same reason, this embedding allowed us to generate vectors of the grammatical gender suffixes themselves (*qzd*, *jxb*). Training was performed on each corpus using 300-dimensional vectors for 10 epochs. To capture morphological information, we included character *n*-grams of 3-6 characters .

Measuring Word-Level Social Gender Bias

A word's gender bias was quantified using the Word Embedding Factual Association Test (WEFAT; Caliskan et al., 2017). For a given target noun, we computed its cosine similarity to two sets of words representing the two social genders: for female, F = {*female, woman, girl, daughter, sister*}; for male, M = {*male, man, boy, son, brother*}. Each noun's bias score is the difference between its mean similarity to female vs. male words, divided by the standard deviation of that noun's individual similarities to all individual words (across both sets):

$$s(w,F,M) = \frac{\text{mean}_{f\in F}\cos(\vec{w},\vec{f}) - \text{mean}_{m\in M}\cos(\vec{w},\vec{m})}{\text{SD}_{x\in F\cup M}\cos(\vec{w},\vec{x})}$$

This normalization scales the association score by the variability of the word's similarities. Positive values indicate a stronger association with female words and negative values indicate a stronger association with male words.

Statistical Analysis

We used two linear mixed-effects models predicting the WEFAT scores of nouns that appeared at least 1000 times in each corpus and did not have an inherent social gender (14,008 nouns). Model 1 was fit to data from all 97 corpora (N = 1,373,610 observations). It included main effects of a noun's grammatical gender (feminine = 1 vs. masculine = -1) and 3 corpus-level variables: extent of grammatical gender, and two variables denoting marking mode (as described above). To test whether different aspects of grammatical gender systems amplify implicit social biases, the model also included interactions between a noun's grammatical gender and each of those corpus-level variables. Covariates included each noun's log frequency and its WEFAT score in the baseline corpus. The latter controlled for bias that was already present in a corpus' content (cf. bias introduced by a grammatical gender system). Random effects included an intercept and random slopes for the corpus-level variables by noun type, and an intercept by corpus. Model 2 was fit to data from the subset of corpora created with a random assignment strategy (N = 1,034,238 observations). It had the same structure as Model 1, but added the feminine-to-masculine ratio predictor as a (corpus-level) fixed effect, and its interaction with a noun's grammatical gender. In both models, continuous predictors were centered and scaled.

Results

In the original corpus, the average WEFAT score of nouns showed a slight masculine-bias (i.e., negative; $M = -0.04$, $SD = 0.68$, $range = [-1.87, 1.89]$). The distribution of WEFAT scores across corpora, broken down by the two grammatical gender classes, are shown in **Figure 2**. Descriptively, across corpora, nouns assigned a feminine grammatical gender had a feminine social bias (i.e., positive WEFAT; $M = 0.16$, $SD = 0.73$, $range = [-1.93, 1.95]$), whereas those assigned a masculine grammatical gender had a masculine social bias ($M = -0.18$, $SD = 0.75$, $range = [-1.92, 1.95]$). Statistical model results across all corpora (Model 1) are in **Table 1**.

Hypothesis 1: Grammatical Gender Introduces Bias. Model 1's intercept was positive ($\beta = 0.093$), indicating an overall bias toward feminine social gender across corpora. The main effect of a noun's grammatical gender was positive ($\beta = 0.009$, $p < .001$), indicating that assigning feminine grammatical gender to a noun increased its association with socially feminine words (e.g., *woman*, *girl*), whereas assigning masculine grammatical gender increased its association with socially masculine words (e.g., *man*, *boy*). Note that this analysis includes some corpora where gender assignment was based on human associations with social gender; for results restricted to corpora with random gender assignment, see Model 2 below.

Hypothesis 2: The Extent of Agreement Amplifies Bias. The interaction between a noun's grammatical gender and the extent of agreement marking in a corpus was significant, and over four times larger than the main effect of grammatical gender class ($\beta = 0.037$, $p < .001$). Thus, bias effects compound

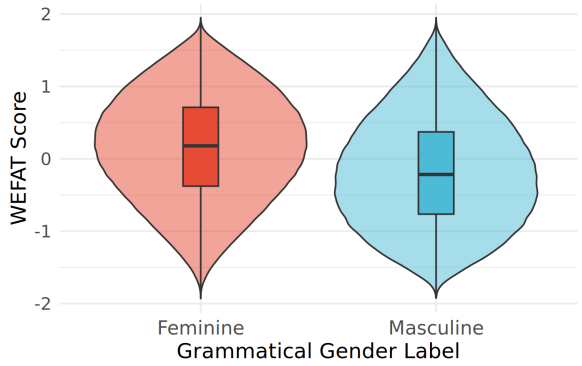

Figure 2: *Violin plots of WEFAT scores by grammatical gender class, with embedded box plots (median + interquantile range). Positive (negative) values indicate stronger associations with socially feminine (masculine) gender.*

Table 1: *Fixed Effects Coefficients for All Corpora (Model 1)*

Predictor	β	*SE*	*t*
Intercept	0.093	0.009	10.37**
GG label	0.009	0.001	14.69**
Extent	0.021	0.005	4.21**
Single marked	−0.157	0.011	−14.88**
Marked gender	−0.125	0.007	−17.93**
GG × Extent	0.037	<.001	109.07**
GG × Single marked	0.228	0.001	320.55**
GG × Marked gender	0.018	<.001	44.37**
GG × Baseline bias	−0.001	<.001	−2.99*
Log frequency	0.028	0.003	10.76**
Baseline bias	0.420	0.003	162.56**

Note. GG = grammatical gender. *$p < .01$. **$p < .001$.

as grammatical gender agreement extends across more parts of speech (**Figure 3**).

Hypothesis 3: Asymmetric Marking Amplifies Bias. The interaction between a noun's grammatical gender and the mode of agreement marking was the strongest effect of interest in the model ($\beta = 0.228$, $p < .001$). Compared to corpora that marked both grammatical genders ($\beta = 0.009$), corpora where only one grammatical gender was explicitly marked had nouns with social biases that were approximately 25 times stronger (i.e., the divergence in WEFAT scores between grammatically-feminine and -masculine nouns was greatly amplified). Additionally, for corpora that marked only one grammatical gender, we found a significant interaction between a noun's grammatical gender and the grammatical gender that was marked in a corpus (feminine vs. masculine) ($\beta = 0.018$, $p < .001$). Specifically, marking only feminine grammatical gender amplified the bias of nouns towards their corresponding social gender more strongly than marking only masculine grammatical gender did.

Baseline Bias. The baseline bias of a noun (i.e., its WEFAT in the original corpus) was the strongest predictor in the model ($\beta = 0.420$, $p < .001$). A noun's social gender association in the absence of any grammatical gender system, which was implicit in its patterns of use in language, strongly predicted its gender association in the grammatically gendered corpora. This baseline bias likely reflects mostly conceptual and cultural meaning, which have a stronger effect on implicit bias compared to grammatical gender. The interaction between a noun's baseline bias and its assigned grammatical gender was small, but significant and negative ($\beta = -0.001$, $p = .003$), suggesting that nouns already strongly biased in the baseline corpus were slightly less affected by grammatical gender manipulations compared to more "neutral" nouns.

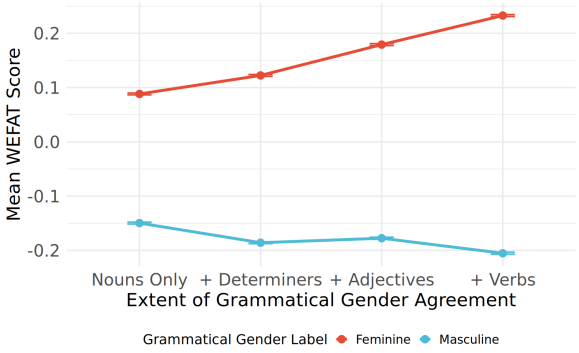

Figure 3: *Mean WEFAT scores by extent of grammatical gender agreement in a corpus (number of marked parts of speech; x-axis) and the grammatical gender of a noun (red vs. blue lines). Error bars show standard errors.*

Model 2: Random Assignment Corpora. Model 2, fitted on the subset of corpora with random gender assignment, showed similar patterns to Model 1. Our hypotheses were again supported: grammatical gender introduced bias ($\beta = 0.009$, $t = 13.25$, $p < .001$), the extent of marking amplified this effect ($\beta = 0.038$, $t = 95.92$, $p < .001$), and asymmetric marking produced the largest amplification ($\beta = 0.230$, $t = 278.83$, $p < .001$). The interaction between a noun's grammatical gender and the ratio of grammatical genders in a corpus was significant ($\beta = 0.005$, $t = 8.19$, $p < .001$): the bias of a noun was stronger when the corpus contained a higher proportion of nouns matching its grammatical gender (e.g., a grammatically feminine noun showed more positive WEFAT scores in corpora where the class of feminine nouns was larger). However, this effect was considerably smaller than those of the other corpus-level variables.

Discussion

We tested whether grammatical gender, a structural property of language, can shape the representation of words to amplify implicit biases related to social gender. Using word embeddings as a (simplified) computational model of lexical representations, we found that introducing grammatical gender altered the organization of words: nouns assigned a feminine grammatical gender became more similar to socially feminine words (e.g., *woman*, *female*), and nouns assigned a masculine grammatical gender—to socially masculine words (e.g., *man*, *male*). This effect occurred despite holding content (conceptual and cultural information) constant across corpora.

Our findings demonstrate that a largely arbitrary grammatical system can "warp" representations of meaning by changing a word's usage patterns. When a noun repeatedly co-occurs

with morphological markings for grammatical gender agreement, this distributional information becomes part of the statistical structure that a word embedding learns and can shift the position of nouns in the embedding space. For example, because grammatically feminine nouns like ***woman****qzd* and ***chair****qzd* co-occur with the same agreement marker, and alongside the same agreement-marked words from other parts of speech (e.g., ***the****qzd*, ***tall****qzd*), they share distributional context and end up in similar regions of the embedding space.

One important finding in this study concerns asymmetry in morphological marking. Systems where only one grammatical gender is explicitly marked, while the other remains unmarked / generic, produced much stronger implicit gender biases in the embedding space compared to systems that marked both genders. Asymmetric marking creates an imbalance between noun classes, because only one class is consistently signaled by additional morphological material, amplifying the influence of social gender information on the structure of the embedding space. We also found that extending grammatical gender agreement across more parts of speech strengthened social gender associations at the level of individual nouns. This result indicates that grammatical gender cues accumulate locally, reinforcing representations consistent with social gender.

Critically, the preexisting association of a noun with a social gender (feminine vs. masculine) in the baseline corpus—likely driven mostly by conceptual and/or cultural semantics—was overall maintained in corpora where grammatical gender was introduced. Thus, grammatical gender does not override or replace patterns already present, but it can amplify them.

The implications of our results are twofold. First, they inform theories of the human mental lexicon insofar as our minds, like word embeddings, track patterns of language use and incorporate such distributional information into lexical representations. Our findings suggest that—independent of cultural differences—languages that differ from one another in grammatical gender may have users who systematically differ in the social biases implicit in their lexicon. These biases may in turn affect how people use words and reason verbally, consistent with evidence that implicit social biases in word embeddings correlate with human implicit associations (Caliskan et al., 2017; Lewis & Lupyan, 2020). More broadly, our results may extend beyond gender and reflect a general property of how statistical learning interacts with morphological structure.

We emphasize that our study characterizes the influence of grammatical gender on representations of words derived from their usage patterns, not on the organization of non-linguistic concepts. In the case of grammatical gender, the way people think and behave when using language can be dissociated from how they think and behave when relying on other modalities (Elpers et al., 2022). This dissociation has been observed in multiple domains, such as when reasoning about event structure (Goldin-Meadow et al., 2008) or object motion (Papafragou and Selimis, 2010; Papafragou et al., 2008; Skordos et al., 2020) (but other domains are strongly influenced by language, e.g., numerical reasoning; Pitt et al., 2022). Still, given the essential role of verbal communication, implicit biases in lexical representations could have behavioral consequences—though the extent to which embedding-level biases translate into human behavior remains an open question.

The second implication of our study concerns language technologies. Large language models and related NLP systems rely on word embeddings as a foundation. Biases encoded in these embeddings can therefore propagate to higher-level model behavior. The present results suggest that bias in such systems may arise not only from biased content, but also from structural properties of the language used to train them. Efforts to de-bias these technologies should take the effect of grammatical gender into account (Zhou et al., 2019).

Methodologically, our study demonstrates the value of artificial corpus manipulations for isolating linguistic effects. Because all 97 corpora we created were identical in content, differing only in their grammatical gender systems, the differences in the social bias implicit in word embeddings can be confidently attributed to grammatical structure itself. This approach complements other methods, based on cross-linguistic comparisons of corpora (e.g., Flint & Ivanova, 2024) or human language users (Cohen et al., 2023; Davis & Reynolds, 2018; Galor et al., 2020; Gygax et al., 2008; Hicks et al., 2014; Jakiela & Ozier, 2020; Kollmayer et al., 2018; Kricheli-Katz & Regev, 2021; Mavisakalyan, 2015; Osypenko et al., 2025; Prewitt-Freilino et al., 2011; Storme & Storme, 2025; van der Velde et al., 2015; Wasserman & Weseley, 2009).

Our approach has several limitations. First, all artificial languages were derived from English, which has minimal grammatical gender (in the form of gendered pronouns). The generated corpora may therefore not fully capture the distributional properties of natural languages with grammatical gender.

Second, we used the same morphological suffix (e.g., *-qzd*) across all parts of speech requiring agreement. In natural languages, agreement morphemes can differ across syntactic categories. Because FastText represents words through character *n*-grams, these suffixes—not only full words—serve as context. The repeated occurrence of a noun near different parts of speech that are all marked with the same suffix might exaggerate the effects of grammatical gender on gender bias.

Third, as noted in the Methods, we aligned grammatical and inherent gender for all nouns with biological sex, whereas natural languages occasionally allow mismatches. Our simulation thus represents a simplified case of gender marking.

Finally, we used static embeddings (FastText), which assign a single vector per word type regardless of context. Contextualized models (e.g., GPT; Radford et al., 2018; BERT; Devlin et al., 2019) may respond differently to grammatical gender.

In conclusion, this study demonstrates that arbitrary grammatical structure, by modulating patterns of language use, can influence how word meaning is represented. Overall, grammatical gender does not introduce new semantic associations, but can systematically amplify existing ones. Hence, our results underscore the role of linguistic structure—not just content—in shaping socially-biased representations.

References

Bartl, M., Nissim, M., & Gatt, A. (2020). Unmasking contextual stereotypes: Measuring and mitigating BERT's gender bias [https://aclanthology.org/2020.gebnlp-1.1/]. *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, 1–16.

Bhardwaj, R., Majumder, N., & Poria, S. (2021). Investigating gender bias in BERT. *Cognitive Computation*, *13*(4), 1008–1018.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *TACL*, *5*, 135–146.

Boleda, G. (2019). Distributional semantics and linguistic theory. *Annual Review of Linguistics*, *6*, 213–234.

Bozdag, M., Sevim, N., & Koç, A. (2024). Measuring and mitigating gender bias in legal contextualized language models. *ACM Transactions on Knowledge Discovery from Data*, *18*(4), 1–26.

Bybee, J. (2010). *Language, usage and cognition*. Cambridge University Press.

Bybee, J. L. (1985). *Morphology: A study of the relation between meaning and form*. John Benjamins Publishing Company.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*, 183–186.

Cohen, A., Karelitz, T., Kricheli-Katz, T., Pumpian, S., & Regev, T. (2023). *Gender-neutral language and gender disparities* (NBER Working Paper No. 31400). National Bureau of Economic Research.

Davis, L., & Reynolds, M. (2018). Gendered language and the educational gender gap. *Economics Letters*, *168*, 46–48.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding [https://aclanthology.org/N19-1423/]. *Proceedings of NAACL-HLT 2019*, 4171–4186.

Elpers, N., Jensen, G., & Holmes, K. J. (2022). Does grammatical gender affect object concepts? Registered replication of Phillips and Boroditsky (2003). *Journal of Memory and Language*, *127*, 104357.

Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955. In *Studies in linguistic analysis*. Blackwell.

Flint, G. R., & Ivanova, A. (2024). Testing a distributional semantics account of grammatical gender effects on semantic gender perception. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *46*.

Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., Ahmed, N. K., et al. (2024). Bias and fairness in large language models: A survey. *Computational Linguistics*, 1–79.

Galor, O., Özak, Ö., & Sarid, A. (2020). Linguistic traits and human capital formation. *AEA Papers and Proceedings*, *110*, 309–313.

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, *115*(16), E3635–E3644.

Goldberg, A. E. (1995). *Constructions: A construction grammar approach to argument structure*. University of Chicago Press.

Goldin-Meadow, S., So, W. C., Özyürek, A., & Mylander, C. (2008). The natural order of events: How speakers of different languages represent events nonverbally. *Proceedings of the National Academy of Sciences*, *105*(27), 9163–9168.

Gygax, P., Gabriel, U., Sarrasin, O., Oakhill, J., & Garnham, A. (2008). Generically intended, but specifically interpreted. *Language and Cognitive Processes*, *23*, 464–485.

Harris, Z. S. (1954). Distributional structure. *Word*, *10*, 146–162.

Hicks, D. L., Santacreu-Vasut, E., & Shoham, A. (2014). Does mother tongue make for women's work? Linguistics, household labor, and gender identity. *Journal of Economic Behavior & Organization*, *110*, 19–44.

Jakiela, P., & Ozier, O. (2020). *Gendered language* (IZA Discussion Paper No. 13568). Institute of Labor Economics (IZA).

Keissar, I., & Meltzer-Asscher, A. (2023). Only fully matching attractors produce attraction: Evidence from Hebrew comprehension. *36th Conference on Human Sentence Processing*.

Kollmayer, M., Pfaffel, A., Schober, B., & Brandt, L. (2018). Breaking away from the male stereotype of a specialist: Gendered language affects performance in a thinking task. *Frontiers in Psychology*, *9*, 985.

Kozlowski, A. C., Taddy, M., & Evans, J. A. (2019). The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review*, *84*(5), 905–949.

Kricheli-Katz, T., & Regev, T. (2021). The effect of language on performance: Do gendered languages fail women in maths? *npj Science of Learning*, *6*, 9.

Kurdi, B., Mann, T. C., Charlesworth, T. E., & Banaji, M. R. (2019). The relationship between implicit intergroup attitudes and beliefs. *Proceedings of the National Academy of Sciences*, *116*(13), 5862–5871.

Kurita, K., Vyas, N., Pareek, A., Black, A. W., & Tsvetkov, Y. (2019). Measuring bias in contextualized word representations [https://aclanthology.org/W19-3823/]. *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, 166–172.

Lake, B. M., & Murphy, G. L. (2021). Word meaning in minds and machines. *Psychological Review*, *130*, 401–431.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*, 1126–1177.

Lewis, M., & Lupyan, G. (2020). Gender stereotypes are reflected in the distributional structure of 25 languages. *Nature Human Behaviour*, *4*, 1021–1028.

Mavisakalyan, A. (2015). Gender in language and gender in employment. *Oxford Development Studies*, *43*, 403–424.

Nangia, N., Vania, C., Bhalerao, R., & Bowman, S. R. (2020). CrowS-pairs: A challenge dataset for measuring social biases in masked language models [https://aclanthology.org/2020.emnlp-main.154/]. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1953–1967.

Osypenko, O., Brandt, S., & Athanasopoulos, P. (2025). Between two grammatical gender systems: Exploring the impact of grammatical gender on memory recall in Ukrainian–Russian simultaneous bilinguals. *Cognitive Science*, *49*.

Panchenko, A., Ruppert, E., Faralli, S., Ponzetto, S. P., & Biemann, C. (2018). Building a web-scale dependency-parsed corpus from CommonCrawl [https://aclanthology.org/L18-1286/]. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Papafragou, A., Hulbert, J., & Trueswell, J. (2008). Does language guide event perception? Evidence from eye movements. *Cognition*, *108*(1), 155–184.

Papafragou, A., & Selimis, S. (2010). Event categorisation and language: A cross-linguistic study of motion. *Language and Cognitive Processes*, *25*(2), 224–260.

Pitt, B., Gibson, E., & Piantadosi, S. T. (2022). Exact number concepts are limited to the verbal count range. *Psychological Science*, *33*(3), 371–381.

Prewitt-Freilino, J. L., Caswell, T. A., & Laakso, E. K. (2011). The gendering of language. *Sex Roles*, *66*, 268–281.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving language understanding by generative pre-training* (Technical report). OpenAI.

Rotaru, A. S., Vigliocco, G., & Frank, S. L. (2018). Modeling the structure and dynamics of semantic processing. *Cognitive Science*, *42*, 2890–2917.

Saussure, F. D. (2011). *Course in general linguistics* [Original work published 1916]. Columbia University Press.

Scott, G. G., Keitel, A., Becirspahic, M., Yao, B., & Sereno, S. C. (2018). The Glasgow norms: Ratings of 5,500 words on nine scales. *Behavior Research Methods*, *51*, 1258–1270.

Shain, C., Meister, C., Pimentel, T., Cotterell, R., & Levy, R. (2024). Large-scale evidence for logarithmic effects of word predictability on reading time. *PNAS*, *121*, e2307876121.

Skordos, D., Bunger, A., Richards, C., Selimis, S., Trueswell, J., & Papafragou, A. (2020). Motion verbs and memory for motion events. *Cognitive Neuropsychology*, *37*(5–6), 254–270.

Storme, B., & Storme, M. (2025). Feminization is more gender-fair than neutralization. *Journal of Language and Social Psychology*.

Thompson, B., Roberts, S. G., & Lupyan, G. (2020). Cultural influences on word meanings revealed through large-scale semantic alignment. *Nature Human Behaviour*, *4*, 1029–1038.

van der Velde, L., Tyrowicz, J., & Siwinska, J. (2015). Language and (the estimates of) the gender wage gap. *Economics Letters*, *136*, 165–170.

Wasserman, B. D., & Weseley, A. J. (2009). ¿Qué? Quoi? Do languages with grammatical gender promote sexist attitudes? *Sex Roles*, *61*, 634–643.

Wittgenstein, L. (2010). *Philosophical investigations* [Original work published 1953]. Wiley-Blackwell.

Zhou, P., Shi, W., Zhao, J., Huang, K.-H., Chen, M., Cotterell, R., & Chang, K.-W. (2019). Examining gender bias in languages with grammatical gender [https://aclanthology.org/D19-1531/]. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 5276–5284.