Investigating the Frequency Distortion of Word Embeddings and Its Impact on Bias Metrics

Francisco Valentini[1,2] **Juan Cruz Sosa**[3] **Diego Fernandez Slezak**[1,3] **Edgar Altszyler**[1,2]

[1]Instituto de Investigación en Ciencias de la Computación, CONICET-UBA, Argentina
[2]Maestría en Data Mining, Universidad de Buenos Aires (UBA), Argentina
[3]Departamento de Computación, FCEyN, UBA, Argentina

fvalentini@dc.uba.ar, juan.cruz.sosa.92@gmail.com, dfslezak@dc.uba.ar, ealtszyler@dc.uba.ar

Abstract

Recent research has shown that static word embeddings can encode words' frequencies. However, little has been studied about this behavior. In the present work, we study how frequency and semantic similarity relate to one another in static word embeddings, and we assess the impact of this relationship on embedding-based bias metrics. We find that Skip-gram, GloVe and FastText embeddings tend to produce higher similarity between high-frequency words than between other frequency combinations. We show that the association between frequency and similarity also appears when words are randomly shuffled, and holds for different hyperparameter settings. This proves that the patterns we find are neither due to real semantic associations nor to specific parameters choices, and are an artifact produced by the word embeddings. To illustrate how frequencies can affect the measurement of biases related to gender, ethnicity, and affluence, we carry out a controlled experiment that shows that biases can even change sign or reverse their order when word frequencies change.[1]

1 Introduction

Static word embeddings have proven to encode semantic information of words and are therefore useful to solve tasks such as synonym selection and analogical reasoning (Mikolov et al., 2013; Levy et al., 2015). More recent contextualized representations have achieved better results (Ethayarajh, 2019), specially in tasks where the local context of words is important (Sezerer and Tekir, 2021). However, static word embeddings are still widely used in computational social science studies that examine global aspects of corpora. For example, embeddings are trained on specific corpora and are used to compute metrics that quantify societal biases and stereotypes that might be present in the text (Garg et al., 2018; Kozlowski et al., 2019; De-Franza et al., 2020; Jones et al., 2020; Lewis and Lupyan, 2020; Charlesworth et al., 2021). Static embeddings are also used in a wide range of applications like topic coherence evaluation (Aletras and Stevenson, 2013), dream theory analysis (Altszyler et al., 2017), literature studies (Diuk et al., 2012), and cognitive science studies (Mota et al., 2022).

Previous research has found static word embeddings appear to be associated with word frequency in various ways: word frequency correlates with embedding norm (Wilson and Schakel, 2015; Arora et al., 2016), the nearest neighbors of the embeddings of medium-frequency English words are more unstable (Hellrich and Hahn, 2016), there are frequency-related differences in the distribution of the inner products between target and context vectors (Mimno and Thompson, 2017), embeddings can accurately predict whether a word is frequent or rare (Schnabel et al., 2015), and the visual inspection of their top principal components suggest they encode frequency (Gong et al., 2018; Mu and Viswanath, 2018). When it comes to using embeddings to measure bias in text, Valentini et al. (2022) found that gender embedding-based bias metrics can spuriously depend on word frequency.

Our work addresses several gaps in the existing literature regarding the frequency distortion of static word embeddings and its impact on the quantification of biases in corpora. Even if it has been pointed out that embeddings can encode frequency, this is the first study that:

1. Comprehensively investigates the association between frequency and similarity in commonly used embeddings.
2. Examines whether embeddings encode frequency due to undesirable properties of embed-

[1]Code for the paper is available at https://github.com/ftvalentini/EmbeddingsFrequency

Findings of the Association for Computational Linguistics: EMNLP 2023, pages 113–126
December 6-10, 2023

dings or actual properties of corpora.
3. Explores the persistence of the frequency distortion under different hyperparameter settings.
4. Assesses the impact on a computational social science application, namely bias measurement.

2 The dependence of word embeddings on frequency

As a first step, we seek to describe the association between word frequency and cosine similarity, commonly used to measure semantic closeness.

We use the 2021 English Wikipedia as a corpus to train embeddings with word2vec with skip-gram with negative-sampling (SGNS, Mikolov et al., 2013), GloVe (Pennington et al., 2014) and FastText (Bojanowski et al., 2017) with default hyperparameters (see details in Appendix A). We group all words by their $\log_{10}$ frequency, and use matrices to represent the mean cosine similarity of 500 randomly sampled pairs of words for each combination of bins, excluding comparisons between the same word.

In the three methods the mean cosine similarity is higher between high frequency words than between any other combination of frequencies (Figure 1). Unlike SGNS and FastText, the GloVe mean similarity is moderate to high between words of the same frequency range (the matrix diagonal) and low between words of different frequencies (e.g. words with 10^6 vs 10^2 frequencies).

These results seem to imply that word frequencies influence how similar two words are, raising the following questions: is this due to an artifact of the embeddings? Or does it reveal actual properties of the corpus; for example, that high-frequency words are actually semantically closer on average to one another than the rest of the vocabulary? We conduct the following study to answer this.

2.1 Experiments

Following Valentini et al. (2022)'s approach, we produce a randomly shuffled Wikipedia corpus, with tokens distributed at random across the text. As co-occurrences are random, words retain their frequency but any contextual information is lost. We train embeddings on this corpus and repeat the analysis from the previous section (as in Figure 1).

If any association is found between similarity and frequency in this setting, it should be explained only by word frequencies. If embeddings don't capture frequency, we would expect a uniform distribution of cosine similarity across all frequency combinations: the similarity of any pair of words should be on average the same.

We find that the mean cosine similarities of embeddings trained on the shuffled corpus depend on the frequencies of the words, and this happens in different ways depending on the method (Figure 2). When comparing frequent words (frequency around 10^4 and above) to one another, all embeddings tend to yield high similarities; and similarity tends to drop in different ways when doing other comparisons.

We obtain the same qualitative result when using an Euclidean distance-based similarity measure: the similarity of any two words depends heavily on their frequencies (Figures 8 and 9 in Appendix B). Therefore the frequency-based effect is not caused by the choice of cosine similarity.

The fact that shuffling words prior to training does not yield a uniform distribution of similarity across frequencies suggests that embeddings tend to encode frequency. To further assess this we do PCA on the vectors of a sample of words stratified by frequency and inspect the centroids of the top two components of each frequency bin (top panel of Figure 3). PCA finds the dimensions with the most variability, and we find that the top two are highly associated with the frequency dimension. Therefore the geometry of vectors trained on the original corpus encodes training data frequencies, which is consistent with the literature's previous findings.

We also run PCA on the vectors trained on the shuffled corpus to confirm that this is not a result of properties of the corpus but rather an artifact of embeddings. The trend is more pronounced in this setting: words with varying frequency tend to live in distinct regions in the embedding space (bottom panel of Figure 3). Thus embedding-based similarity metrics can detect semantic closeness even when there shouldn't be any. Four additional independent random shuffles of the corpus yielded the same qualitative results.

PCA in the shuffled corpus also reveals that the vectors of low frequency words are more spread out and therefore have lower similarity between themselves (as seen in the SGNS heatmap of Figure 2). This might occur because the distribution of co-occurrences of low-frequency words is less varied, and thus noisier, even when shuffling the corpus. On the other hand, the co-occurrences of higher

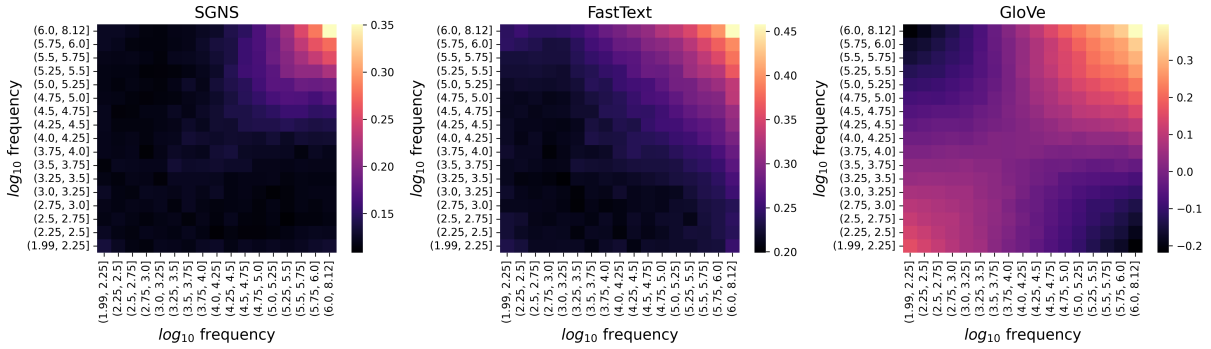

Figure 1: Mean cosine similarity between 500 random word pairs for each combination of frequencies, in embeddings trained on Wikipedia. **Cosine similarity tends to be higher between high frequency words.**

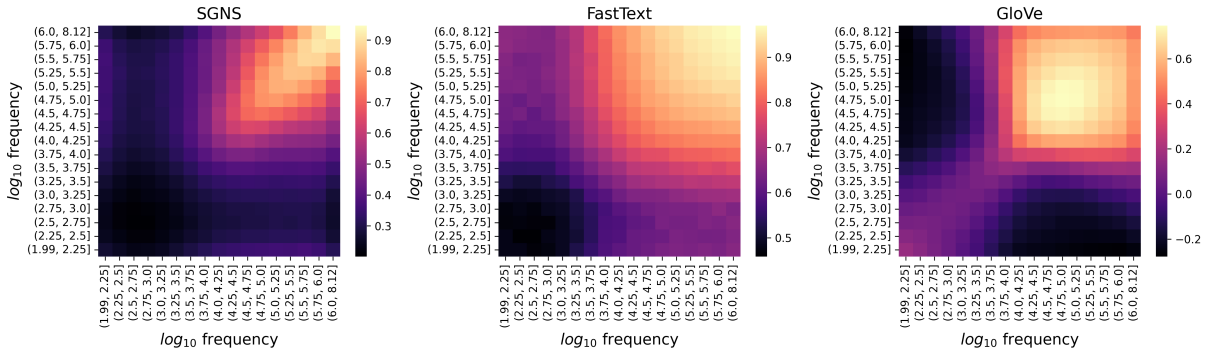

Figure 2: Mean cosine similarity between 500 random word pairs for each combination of frequencies, in embeddings trained on a shuffled version of Wikipedia. **The similarity of embeddings trained on a corpus with random co-occurrences depends on the words' frequencies.**

frequency words are more broadly distributed, and as they are completely random because of the shuffling, the vectors of these words are affected almost exclusively by the frequency dimension, and thus the high similarity between them.

We provide an intuition on how this phenomenon arises during training by visualizing the similarity heatmaps of SGNS embeddings of the shuffled corpus in each epoch (Figure 4). By the end of the first epoch, the vectors of low frequency words have probably moved very little and remain close to their initialization–thus the relatively high similarity between them. In contrast, the vectors of more frequent words are updated with higher variability as their random co-occurrences with other words are more broadly distributed. As training progresses, the frequency dimension becomes more salient and the importance of co-occurrences decreases, so much so that by the last epoch frequency is the dimension that captures the most variance in the vectors (PCA in Figure 3).

2.2 Sensitivity to hyperparameters

We evaluate the robustness of the findings from section 2.1 to embeddings' hyperparameters choices. Following Levy et al. (2015), the hyperparameters values we explore in all methods are:

- Window size (`win`): 2, 5, 10
- Adding context vectors (`w+c`): *yes*, *no*

For SGNS and FastText we also explore the following hyperparameters:

- Context distribution smoothing (`cds`): 0.75, 1
- Number of negative samples (`neg`): 1, 5, 15

Trying all combinations of these hyperparameters results in 6 GloVe, 36 SGNS and 36 FastText settings. With each setting we train embeddings on the shuffled corpus.

In order to measure the association between word frequencies and similarity between embeddings, we compute the root mean squared error (RMSE) between the values of each cell of the similarity heatmap and the overall average. The overall average represents the mean similarity we would

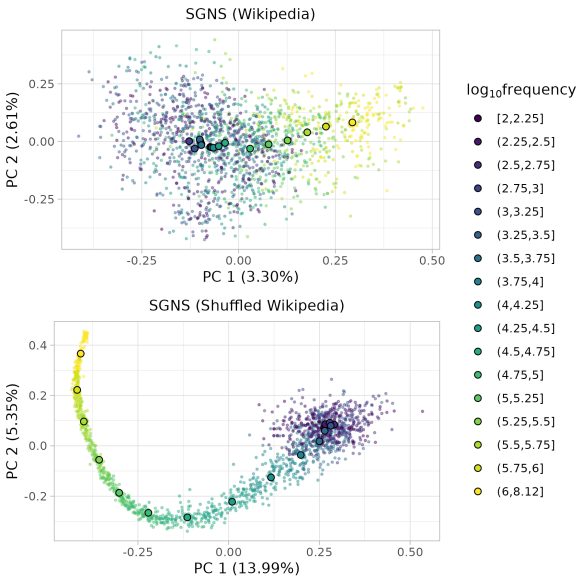

Figure 3: PCA of a sample of SGNS embeddings stratified by frequency, trained on Wikipedia (top) and shuffled Wikipedia (bottom). Vectors are normalized to unit length before PCA and there are 100 words by frequency bin. Centroids are displayed with larger markers. Figure 11 in Appendix C displays the plots for GloVe and FastText. **The top two components are highly associated with the frequency dimension: the geometry of vectors encodes the training corpus frequencies.**

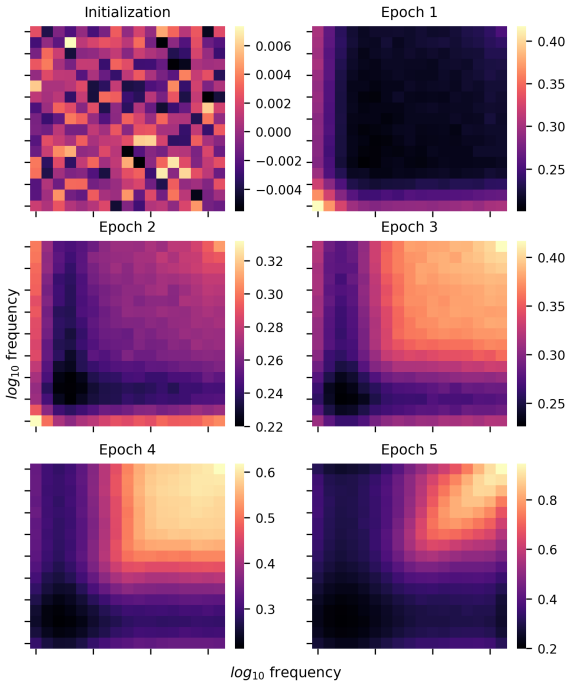

Figure 4: Mean cosine similarity by epoch between 500 random word pairs in SGNS embeddings trained on shuffled Wikipedia. Frequency bins are the same as in Figure 2. **As training progresses, the frequency dimension becomes more salient and the importance of co-occurrences decreases.**

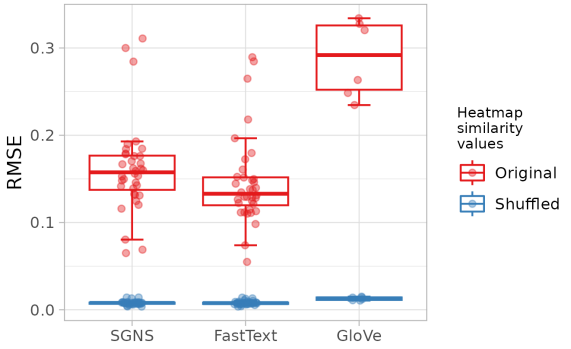

Figure 5: Association between frequency and cosine similarity as measured by the RMSE metric. Each point is a hyperparameter setting of embeddings trained on shuffled Wikipedia. Comparing the RMSE computed with the actual heatmaps (red) with the RMSE computed with heatmaps obtained by shuffling the similarity values of the pairs of words (blue) shows that the metric yields low values when there is no relationship between similarity and frequency. **Word similarity depends on word frequency in all hyperparameter settings.**

see in each cell of the heatmap if there were no association between frequencies and similarity. The larger the RMSE, the larger the deviation from the uniform distribution of similarity across frequency combinations, and thus the stronger the association between frequencies and similarity.

As a control, we compare the actual distribution of RMSE with the distribution of RMSE if there was no frequency effect. We compute this by randomly shuffling the similarity values of the pairs of words used to build the heatmap in each hyperparameter setting. The RMSEs would be close to zero in this case, as the random shuffling would result in a uniform distribution of similarity values across frequencies i.e. no association between frequencies and similarity.

We compute the RMSE metric for cosine similarity in each hyperparameter setting (Figure 5). The RMSE metric is high in all the hyperparameter settings, as compared to the distribution of RMSE obtained by shuffling the similarity values of the pairs of words. This proves that the finding that word similarity depends on word frequencies is robust to hyperparameter choices. In Appendix B we show this is also the case for an Euclidean distance-based similarity measure.

A linear regression analysis that includes the hyperparameter values as predictors and the RMSE as the outcome variable shows that no specific hyperparameter tends to systematically yield higher

or lower RMSE values when using cosine similarity. Only adding context vectors (`w+c` = *yes*) is significantly associated with a greater association between frequency and Euclidean distance (p-value < 0.0001) in all three methods. Refer to Appendix D for more details.

3 Assessing the impact

In computational social science, static word embeddings are typically used to measure societal biases and stereotypes potentially present in corpora. Here we study the effect of the dependence of embeddings on frequency on this type of studies. We highlight that this is different from measuring or mitigating biases in NLP models. Our goal is to assess how much the individual frequency of words might distort the estimates of biases in specific corpora when using word embeddings.

To measure the bias of a target word x we use the difference between the mean similarity of words of context groups A and B with respect to x:

$$\text{Bias}_{\text{WE}} = \underset{a \in A}{\text{mean}} \cos(w_x, w_a) - \underset{b \in B}{\text{mean}} \cos(w_x, w_b) \quad (1)$$

where w_i is the embedding of word i and $\cos(w_i, w_j)$ is the cosine similarity. A and B are set based on the bias to be measured. For instance, to quantify binary gender bias (female/male), gendered nouns and pronouns are used.

Here we use the same bias metric as in Lewis and Lupyan (2020) and Valentini et al. (2022) for its simplicity. Other similar metrics have been used (Bolukbasi et al., 2016; Garg et al., 2018; Kozlowski et al., 2019; Jones et al., 2020) and have shown to yield similar results (Garg et al., 2018).

3.1 Experimental setup

To measure the sensitivity of embedding-based bias to changes in the context words frequencies, we first use the female/male gender bias as a widely studied test case (Garg et al., 2018; Kozlowski et al., 2019; DeFranza et al., 2020; Jones et al., 2020; Lewis and Lupyan, 2020; Charlesworth et al., 2021).

We measure gender bias with equation 1 with $A = \{she\}$ and $B = \{he\}$ (Bolukbasi et al., 2016). We seek to train embeddings on corpora where one of the words (A) has the original corpus' frequency, while the other word (B) has a target frequency level. To achieve this, we randomly drop sentences containing B until the target frequency is reached. We set three target frequencies for B: 10^4, 10^5, and 10^6. This implies creating three resampled corpora (see Table 1).

	she (A)	*he* (B)
Original Wiki.	$10^{6.55}$	$10^{7.07}$
Undersampled Wiki. 1	$10^{6.53}$	10^{6}
Undersampled Wiki. 2	$10^{6.52}$	10^{5}
Undersampled Wiki. 3	$10^{6.52}$	10^{4}

Table 1: Frequency of context words in the undersampling experiment that drops sentences with word B (*he*). The frequency of A (*she*) decreases but to a minor extent.

Using only one word in each context group to measure bias (as in Bolukbasi et al., 2016) allows us to ascribe any shifts in bias to the change in the frequency of one of the words. This simplifies the experiment and the conclusions we can draw from it, as compared to the case where multiple words are used in each context group.

With the embeddings trained on the resampled corpora and on the original corpus, we compute equation 1 on the words from the Glasgow Norms, a set of around 5,500 words with a score of gender association as perceived by human judgment (Scott et al., 2019). See details in Appendix E.

Our hypothesis is that bias can be heavily affected by the frequencies of the context words, so much so that answers to questions of the type "*what are the most gender-biased words in this corpus?*" can be highly dependent on the frequencies of words in the corpus being studied.

3.2 Results

The effect of frequencies on gender Bias_{WE} is different in each method (leftmost panel in Figure 6). SGNS has an issue when measuring bias in high frequency words: in these words the association between Bias_{WE} and the frequency of B is negative, while it is approximately constant for the rest of the frequency ranges.

In FastText embeddings this negative association is observed across all frequency bins. Moreover, when the frequency of *he* is low enough, all words have a positive (female) bias. This means the bias of specific words might appear to be high, when in fact the average bias of *all* words is high.

Context words frequencies have the most influence on bias when using GloVe, as there are very different effects in each frequency bin. Both the

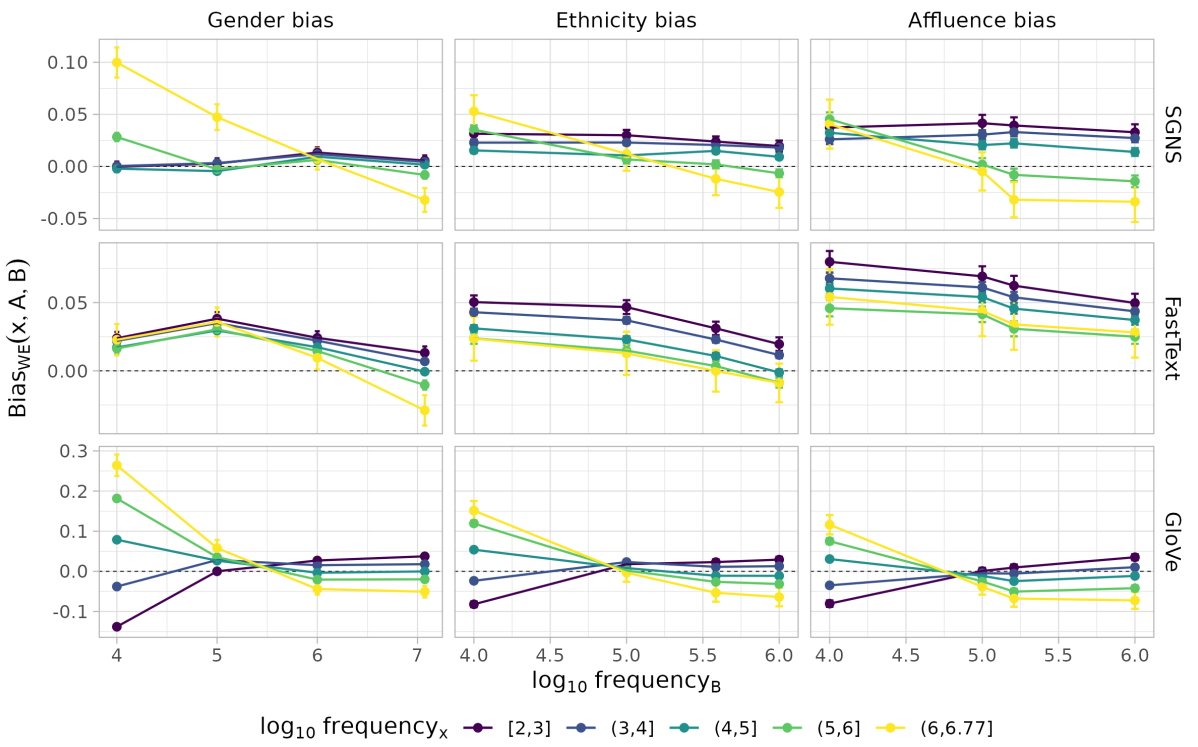

Figure 6: Mean biases of selected words grouped by frequency. Horizontal axes represent corpora where the frequency of B varies and the frequency of A is almost constant (see Tables 1, 2 and 3). Averages are plotted with bootstrap confidence intervals. Gender, ethnicity, and affluence biases are measured in 4,384, 4,641, and 4,617 target words, respectively. **Changes in word frequencies can generate substantial changes in estimates of gender, ethnicity and affluence bias, even if the underlying distribution of co-occurrences remains constant.**

level and ranking of bias are highly affected by frequency. More frequent target words tend to stick to the more frequent context word and less frequent words are attracted to the less frequent context.

In summary, gender bias estimates can change substantially with the three methods even if the underlying distribution of co-occurrences remains constant. These shifts are triggered by the change in the context words' frequencies, which is an undesirable property in similarity measurements. In Figure 12 from Appendix E we show that the frequency dependence persists when undersampling word A instead of B.

3.2.1 Qualitative analysis of individual words

To illustrate how this property can lead to misleading conclusions, we perform a qualitative analysis of the Bias_{WE} of individual words with GloVe, which seems to have the strongest frequency-based distortion.

We classify words by their perceived genderedness according to human judgment: "male" if their Glasgow gender norm is equal to or less than 2, "female" if it is 6 or higher, and "neutral" otherwise. For each class of words we sample a word for each one of five frequency bins and we study the changes in bias according to the frequency of *he* in the corpus (Figure 7).

As observed in Figure 6, the Bias_{WE} of frequent words (frequency 10^5 onwards) is inversely correlated to the frequency of *he*, while the association is positive for less frequent words (below 10^4 occurrences). This occurs regardless of whether the words are perceived to be "male", "female" or "neutral".

The bias is so dependent on the frequencies of the context words that the ranking of words can be reverted. For instance, if the frequency of *he* is low enough, we might end up believing that "male-associated" words like *war* and *battle*, or "neutral" words as *new* or *art*, are female-biased words in our corpus, with even larger values than *hostess*. At the same time, *lioness* might tend to appear as male-biased, even more than *wrestler*, *battle* or *war*.

What is more, this behavior holds both for words that are inherently gendered, e.g. *daughter* or *grandpa*, and for words that are stereotypically

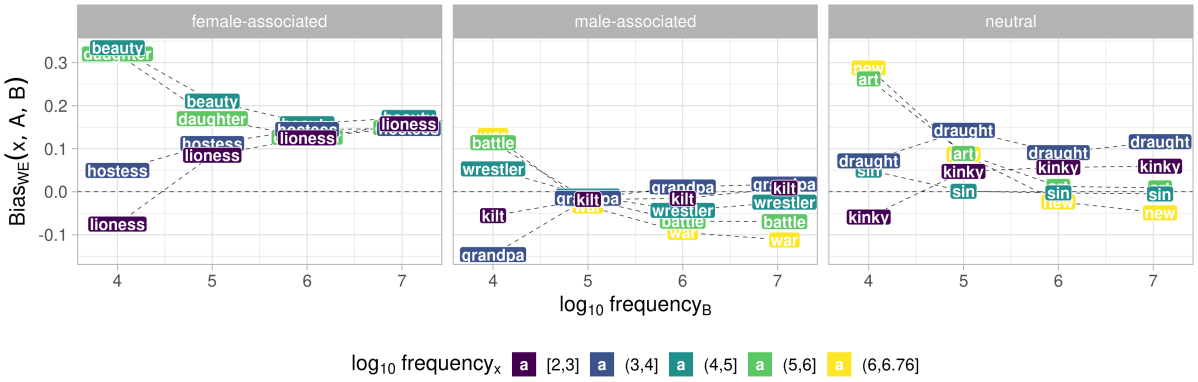

Figure 7: GloVe gender bias. Selected words are classified by frequency bin and by their perceived genderedness by human judgment ("female-associated", "male-associated", and "neutral"). Horizontal axes represent corpora where the frequency of B (*he*) varies and the frequency of A (*she*) is almost constant (see Table 1). **Biases can even change sign or ranking when frequencies change.**

associated with gender, e.g. *beauty* or *battle* (refer to Figure 13 in Appendix E.1 for a more detailed analysis). In Appendix E.1 we also replicate the analysis of specific words for SGNS and FastText.

3.2.2 Other biases

To show that the findings do not apply to gender bias alone, we assess the impact on the measurement of bias along other cultural dimensions, namely ethnicity (Caliskan et al., 2017; Kozlowski et al., 2019) and affluence (Kozlowski et al., 2019).

We measure ethnicity bias with $A = \{african\}$ and $B = \{european\}$, and affluence bias with $A = \{rich\}$ and $B = \{poor\}$. The words were chosen from Kozlowski et al. (2019) and aiming at reducing the ambiguity of the association being measured (e.g. african/european is less ambiguous than black/white).

We follow the same approach from section 3.1. In these cases achieving the desired frequency levels implies also oversampling i.e. randomly replicating sentences containing B (see Tables 2 and 3). As with gender bias, we use the words from the Glasgow Norms as target words, applying the same filtering described in Appendix E.

The findings are qualitatively the same as those in section 3.2: the estimates of bias can be highly dependent on the frequencies of the words involved, and the effect is different in each set of embeddings (middle and right panels of Figure 6). Moreover, the affluence and ethnicity biases of each frequency range vary in a very similar manner to gender bias. This further supports the claim that the frequency of the words being compared heavily affects similarity scores and bias estimates, independently of the specific context words we have chosen to use in these experiments.

	african (A)	*european* (B)
Oversampled Wiki.	$10^{5.39}$	10^{6}
Original Wiki.	$10^{5.37}$	$10^{5.58}$
Undersampled Wiki. 1	$10^{5.36}$	10^{5}
Undersampled Wiki. 2	$10^{5.36}$	10^{4}

Table 2: Frequency of context words in the resampling experiment that either drops or replicates sentences with B (*european*). The frequency of word A (*african*) changes but to a minor extent.

	rich (A)	*poor* (B)
Oversampled Wiki.	$10^{5.05}$	10^{6}
Original Wiki.	$10^{4.95}$	$10^{5.21}$
Undersampled Wiki. 1	$10^{4.94}$	10^{5}
Undersampled Wiki. 2	$10^{4.93}$	10^{4}

Table 3: Frequency of context words in the resampling experiment that either drops or replicates sentences with B (*poor*). The frequency of word A (*rich*) changes but to a minor extent.

4 Discussion and conclusions

Static word embeddings are useful for computing semantic similarity between words since they capture the semantics of words. To assess biases in texts, computational social scientists frequently use word embedding similarity as a metric.

Static word embeddings can encode information on word frequency, according to earlier research. We examine the relationship between frequency and semantic similarity in SGNS, FastText, and

GloVe embeddings in greater detail in this work. We find that the frequency of the words being compared affects their similarity score. This dependence is also present when words in the training corpus are randomly shuffled, demonstrating that the behavior is an artifact of the embeddings and not a result of actual associations found in the text. Moreover, we find that this frequency distortion persists under different hyperparameter settings.

In computational social science applications like bias measurement, the propensity of embeddings to encode frequency hampers their ability to measure semantic closeness. We conduct a controlled experiment that illustrates how measuring gender, affluence or ethnicity biases using embedding-based metrics might produce inaccurate results. The results indicate that the frequency of words can have a significant impact on the answers to questions like "*what are the most gender biased terms?*" as biases can change sign or ranking when word frequencies are changed.

A way to mitigate the frequency distortion in embedding-based bias metrics could involve creating context groups A and B with words that have similar average frequencies, if possible. When doing this, frequency does not have a systematic effect in the sign of the subtraction of cosine similarities. Another approach is to randomly replicate documents prior to training embeddings so that the frequencies of A and B are balanced. However, it is worth noting that when measuring bias with context groups A and B with multiple words each, it can be challenging to achieve balanced frequencies for all words simultaneously.

Limitations

Experiments were conducted solely on the English language. This means that our findings may not be directly applicable to languages that have more complex morphological features or richer grammatical genders.

Even if our analyses are focused exclusively on the English Wikipedia corpus, we consider that the random-shuffling experiment is sufficiently generic to prove that the dependence on frequency would continue to hold true in other domains.

The embedding-based metric we use to measure biases imply a binary understanding of stereotypes, which excludes other views. The context words were chosen from past studies and aiming at reducing the ambiguity of the associations being measured.

References

Nikolaos Aletras and Mark Stevenson. 2013. Evaluating topic coherence using distributional semantics. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)–Long Papers*, pages 13–22.

Edgar Altszyler, Sidarta Ribeiro, Mariano Sigman, and Diego Fernández Slezak. 2017. The interpretation of dream meaning: Resolving ambiguity using latent semantic analysis in a small corpus of text. *Consciousness and Cognition*, 56:178–187.

Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. 2016. A latent variable model approach to PMI-based word embeddings. *Transactions of the Association for Computational Linguistics*, 4:385–399.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Tessa ES Charlesworth, Victor Yang, Thomas C Mann, Benedek Kurdi, and Mahzarin R Banaji. 2021. Gender stereotypes in natural language: Word embeddings show robust consistency across child and adult language corpora of more than 65 million words. *Psychological Science*, 32(2):218–240.

David DeFranza, Himanshu Mishra, and Arul Mishra. 2020. How language shapes prejudice against women: An examination across 45 world languages. *Journal of Personality and Social Psychology*, 119(1):7–22.

Carlos G Diuk, D Fernandez Slezak, Iván Raskovsky, Mariano Sigman, and Guillermo A Cecchi. 2012. A quantitative philology of introspection. *Frontiers in integrative neuroscience*, 6:80.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, Hong Kong, China. Association for Computational Linguistics.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Chengyue Gong, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tie-Yan Liu. 2018. Frage: Frequency-agnostic word representation. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Johannes Hellrich and Udo Hahn. 2016. Bad Company—Neighborhoods in neural embedding spaces considered harmful. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2785–2796, Osaka, Japan. The COLING 2016 Organizing Committee.

Jason J Jones, Mohammad Ruhul Amin, Jessica Kim, and Steven Skiena. 2020. Stereotypical gender associations in language have decreased over time. *Sociological Science*, 7:1–35.

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

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

Molly Lewis and Gary Lupyan. 2020. Gender stereotypes are reflected in the distributional structure of 25 languages. *Nature Human Behaviour*, 4(10):1021–1028.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

David Mimno and Laure Thompson. 2017. The strange geometry of skip-gram with negative sampling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2873–2878, Copenhagen, Denmark. Association for Computational Linguistics.

Natália Bezerra Mota, Ernesto Soares, Edgar Altszyler, Ignacio Sánchez-Gendriz, Vincenzo Muto, Dominik Heib, Diego F Slezak, Mariano Sigman, Mauro Copelli, Manuel Schabus, et al. 2022. Imagetic and affective measures of memory reverberation diverge at sleep onset in association with theta rhythm. *NeuroImage*, page 119690.

Jiaqi Mu and Pramod Viswanath. 2018. All-but-the-top: Simple and effective postprocessing for word representations. In *International Conference on Learning Representations*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. `http://is.muni.cz/publication/884893/en`.

Tobias Schnabel, Igor Labutov, David Mimno, and Thorsten Joachims. 2015. Evaluation methods for unsupervised word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 298–307, Lisbon, Portugal. Association for Computational Linguistics.

Graham G Scott, Anne Keitel, Marc Becirspahic, Bo Yao, and Sara C Sereno. 2019. The Glasgow Norms: Ratings of 5,500 words on nine scales. *Behavior Research Methods*, 51:1258–1270.

Erhan Sezerer and Selma Tekir. 2021. A survey on neural word embeddings. *arXiv preprint*, arXiv:2110.01804.

Francisco Valentini, Germán Rosati, Diego Fernandez Slezak, and Edgar Altszyler. 2022. The undesirable dependence on frequency of gender bias metrics based on word embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2022*. Association for Computational Linguistics.

Benjamin J Wilson and Adriaan MJ Schakel. 2015. Controlled experiments for word embeddings. *arXiv preprint arXiv:1510.02675*.

A Data and methods

We build the Wikipedia corpus from the April 2021 English Wikipedia dump (`https://archive.org/download/enwiki-20210401`, license CC BY-SA 3.0). Wikipedia is freely available, easily accessible and has been used in previous experiments (Levy et al., 2015). We remove articles with less than 50 words. Pre-processing includes sentence splitting, lowercasing and removing non alpha-numeric symbols, and produces a corpus of 78 million sentences and 1.2 billion tokens.

We train word embeddings with 300 dimensions. All words with less than 100 occurrences are removed before obtaining word-context pairs and we use a sliding window size of 10 tokens by default. SGNS and FastText are trained with Gensim's implementation (Řehůřek and Sojka, 2010, v4.2.0, licensed under GNU LGPLv2.1), and GloVe is trained with Pennington et al. (2014)'s implementation (v1.2, Apache License Version 2.0).

We used a desktop computer with 16 cores Intel Core i7-11700 CPU and 32GB RAM. Depending on the corpus being used, training took between 0.5 and 1.2 hours per epoch with SGNS and FastText, and 5 and 15 minutes per iteration with GloVe.

B Euclidean distance-based similarity metric

Figures 8 and 9 show the association between similarity and word frequency when using a similarity measure based on Euclidean distance.

Figure 10 shows the distribution of the RMSE metric across different hyperparameter settings when using the Euclidean-based measure instead of cosine similarity. The relatively high values of RMSE in red as compared to the baseline values in blue shows that the frequency-based distortion is present in all hyperparameter settings even when using Euclidean distance.

C Principal Components Analysis

Figure 11 shows the top two principal components of each frequency bin in unshuffled Wikipedia and shuffled Wikipedia for embeddings trained with FastText and GloVe.

D Sensitivity to hyperparameters

	SGNS+FT+GloVe		SGNS+FT	
	Cos.	Eucl.	Cos.	Eucl.
`win=5`	0.01	0.03	0.01	0.04
`win=10`	0.02	0.10	0.02	0.13
`w+c=yes`	0.01	0.26***	0.01	0.24***
`neg=5`			−0.01	0.03
`neg=15`			−0.03	0.04
`cds=1`			−0.00	−0.04
N	156	156	144	144

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 4: Coefficients of the linear regressions between the RMSE metric and hyperparameter choices, measuring embedding similarity with both cosine ("Cos.") and negative euclidean distance ("Eucl."). **No hyperparameter choice significantly affects the association between frequency and cosine similarity.**

Table 4 displays the coefficients of a linear model with the hyperparameter values as predictors and the RMSE as the outcome variable, considering the cosine and the Euclidean-based similarity metrics. The only hyperparameter choice that tends to systematically yield higher RMSE values (stronger association between frequency and similarity) is adding context vectors (`w+c` = *yes*), but only when using negative Euclidean distance as similarity metric.

E Impact on bias measurement

The Glasgow Norms (Scott et al., 2019) comprise a set of 5,553 English words rated by subjects who were asked to measure the degree to which each word is associated with male or female behavior on a scale from 1 (feminine) to 7 (masculine). We flip the scale so that the norm represents femaleness according to human judgment.

We discard the norms of homonyms and of words with uppercase characters. Moreover, we only consider words that are in the vocabularies of all embeddings trained on the original corpus and the resampled corpora. Finally, we drop any words that change their frequency bin between corpora. This results in a set of 4,384, 4,641, and 4,617 words to measure gender, ethnicity and affluence bias in Figure 6.

Figure 12 shows the effect of undersampling word A (*she*) instead of B (*he*). The frequencies employed in this experiment are in Table 5. In the same manner as in section 3.2, GloVe exhibits the highest frequency-based distortion, as more frequent target words stick to the more frequent context word (here, *he*) and less frequent words are attracted to the less frequent context (*she*). SGNS also presents the same effect in high frequency words as the one observed in Figure 6. The main difference with respect to the experiment in section 3.2 occurs with FastText. We have no hypothesis about the reason for this discrepancy.

	she (A)	*he* (B)
Original Wiki.	$10^{6.55}$	$10^{7.07}$
Undersampled Wiki. 1	10^{6}	$10^{7.07}$
Undersampled Wiki. 2	10^{5}	$10^{7.07}$
Undersampled Wiki. 3	10^{4}	$10^{7.07}$

Table 5: Frequency of context words in the undersampling experiment that drops sentences with word A (*she*). The frequency of B (*he*) decreases but to a minor extent.

E.1 Qualitative analysis of individual words (SGNS and FastText)

We chose male and female-associated words that are either inherently gendered or stereotypically associated with gender and studied the behavior of

bias in the experiment that undersamples *he* (Figure 13). Results reveal that the frequency distortion affects the bias of both types of words in the same way.

In Figures 14 and 15 we replicate the analysis of specific words for SGNS and FastText, respectively. These words are a subset of the words used to make the leftmost panel of Figure 6.

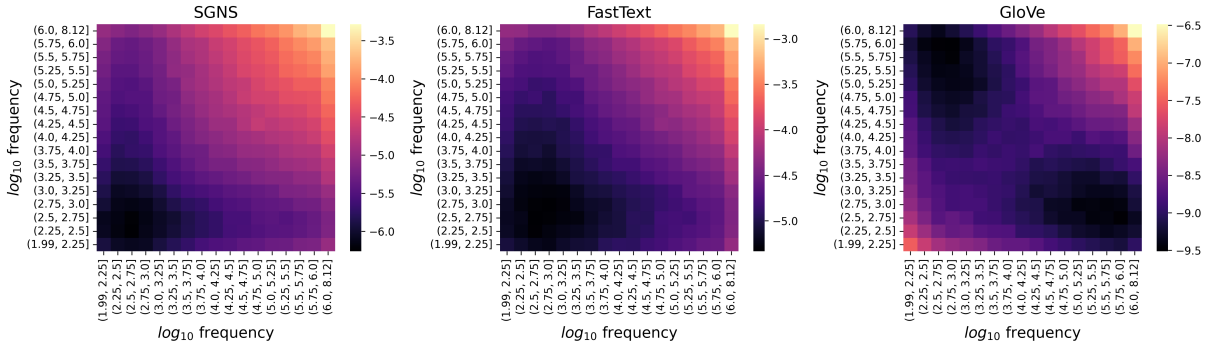

Figure 8: Mean negative Euclidean distance between 500 random word pairs for each combination of frequencies in embeddings trained on Wikipedia.

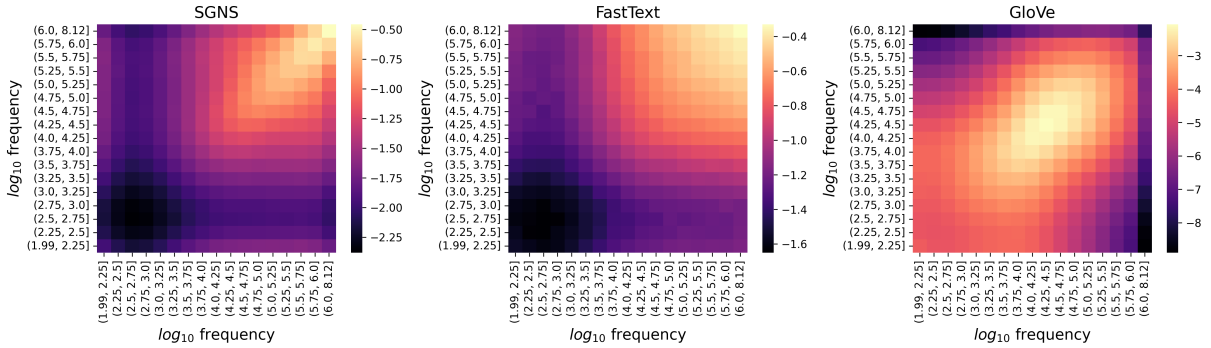

Figure 9: Mean negative Euclidean distance between 500 random word pairs for each combination of frequencies in embeddings trained on a shuffled version of Wikipedia.

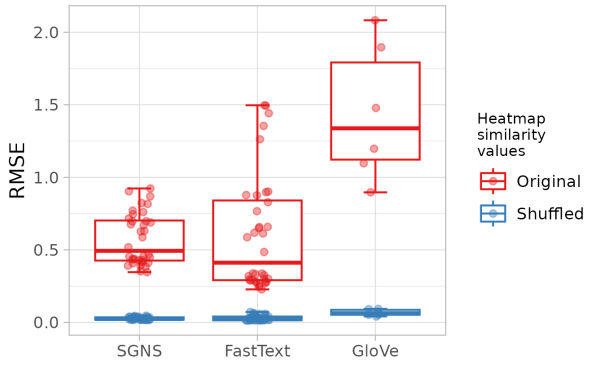

Figure 10: Association between frequency and negative Euclidean distance as measured by the RMSE metric. Each point is a hyperparameter setting of embeddings trained on shuffled Wikipedia. The RMSE computed with the actual heatmaps (red) is compared to the RMSE computed with the heatmaps obtained by shuffling the similarity values of the pairs of words (blue).

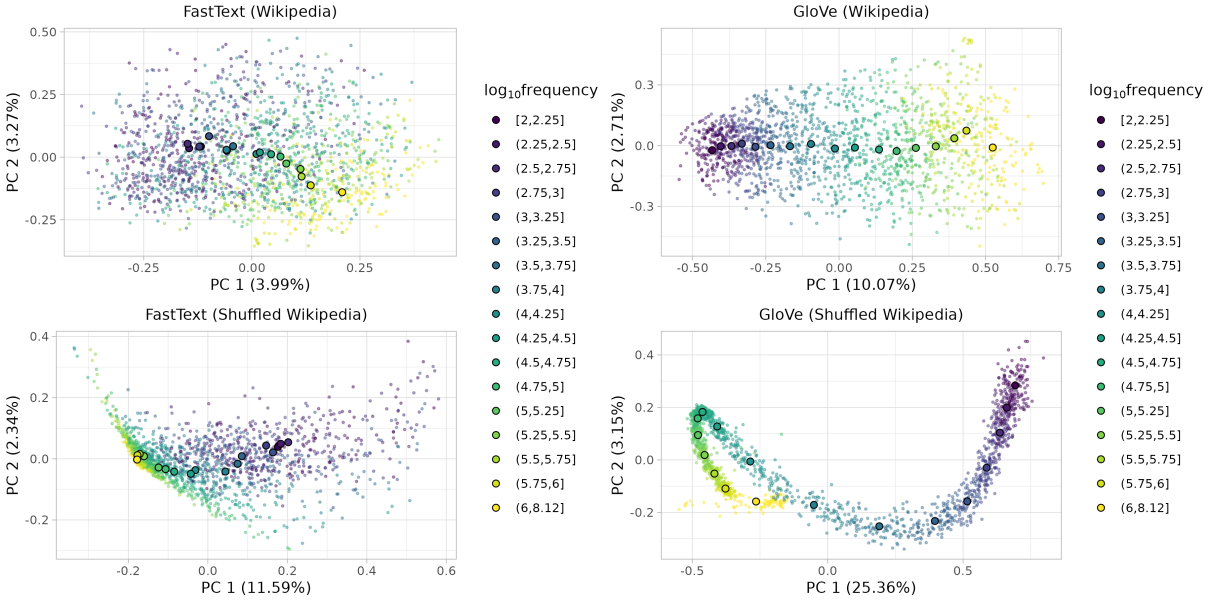

Figure 11: Top principal components of a sample of embeddings stratified by frequency trained on the original Wikipedia (top) and the shuffled Wikipedia (bottom), for FastText (left) and GloVe (right). Vectors are normalized to unit length before PCA and there are 100 words by frequency bin. Centroids are displayed with larger markers.

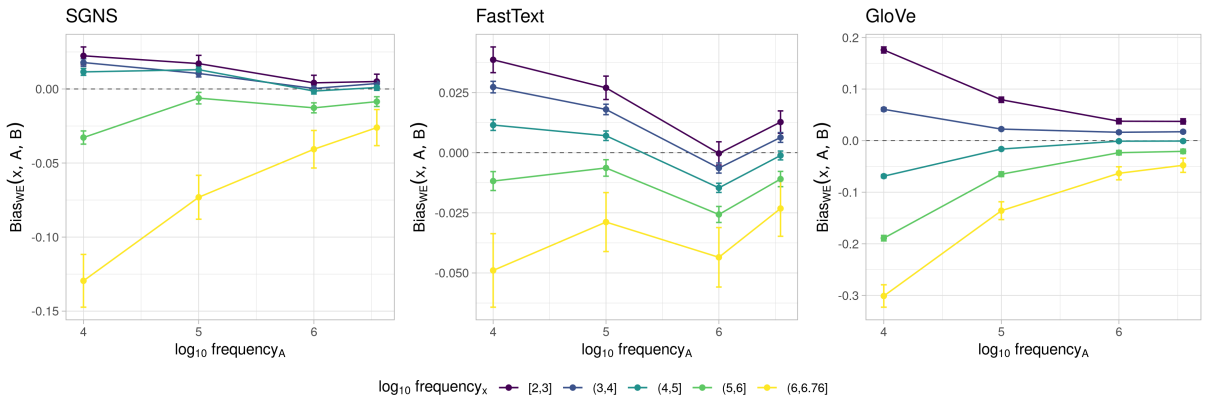

Figure 12: Female/male gender bias of 4,384 target words grouped by frequency. Horizontal axes represent corpora where the frequency of A (*she*) varies and the frequency of B (*he*) is almost constant (see Table 5). Mean bias is plotted with bootstrap confidence intervals.

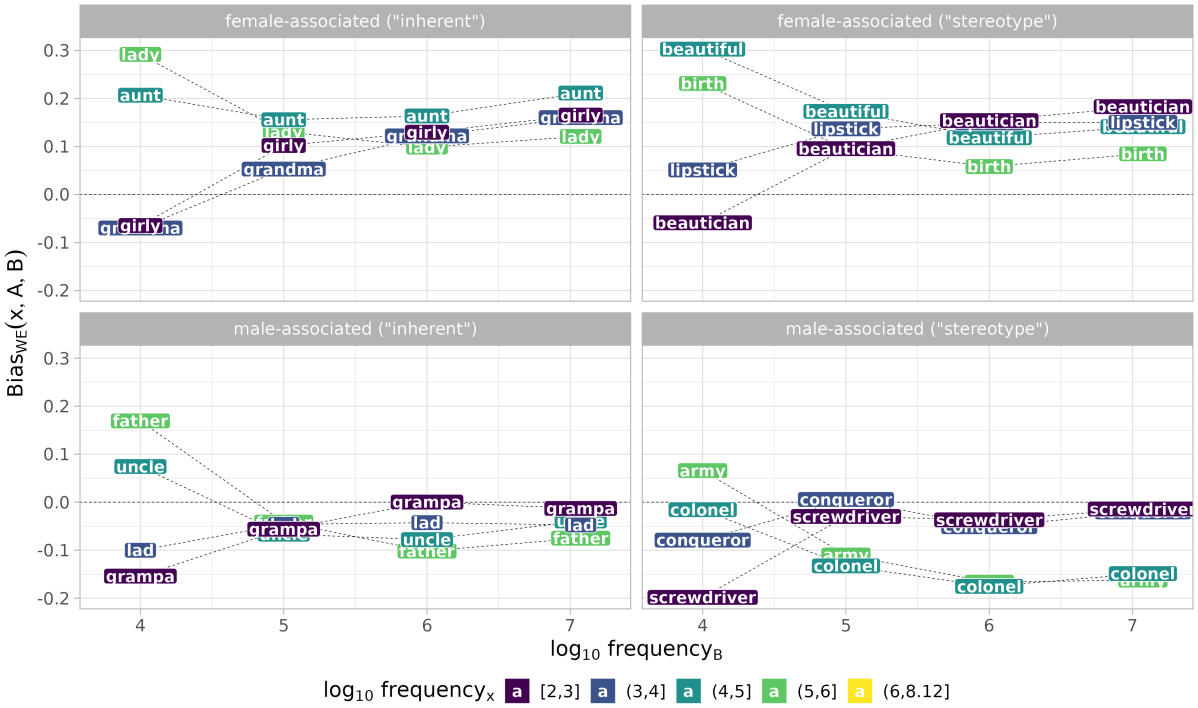

Figure 13: GloVe female/male gender bias of inherently and stereotypically gendered words in the experiment that undersamples B (*he*).

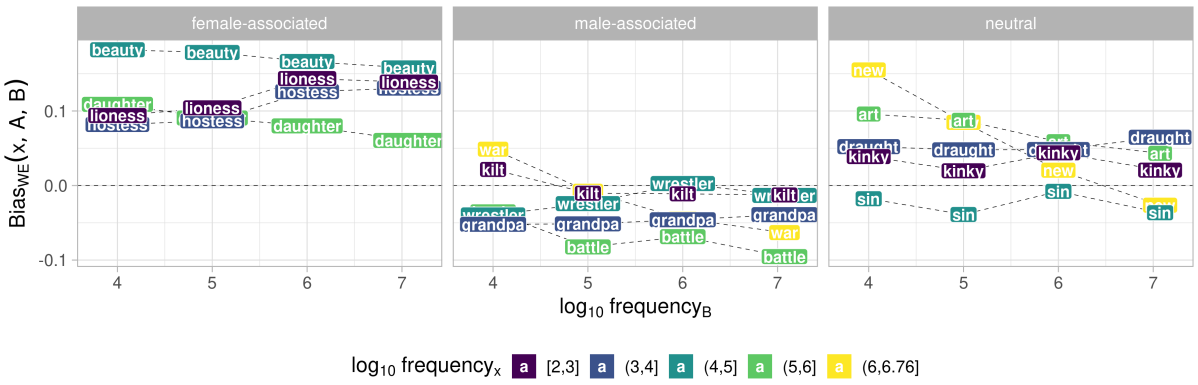

Figure 14: SGNS female/male gender bias of words in the experiment that undersamples B (*he*).

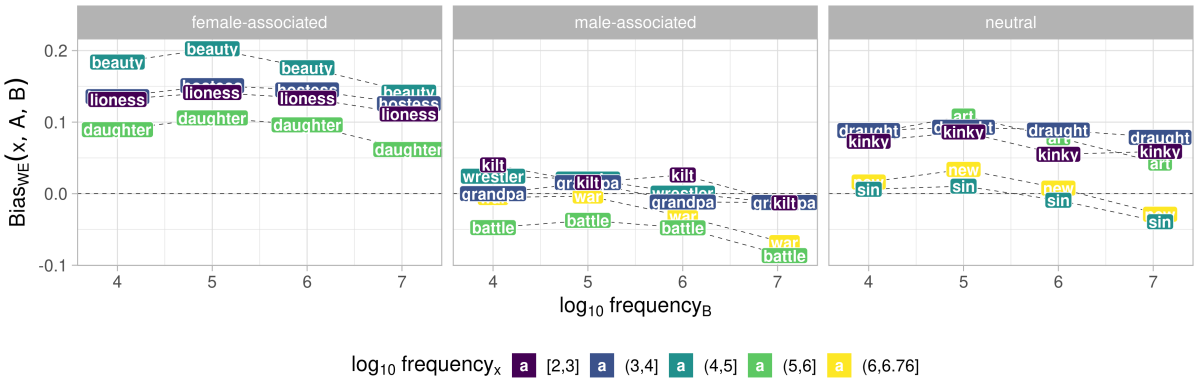

Figure 15: FastText female/male gender bias of words in the experiment that undersamples B (*he*).