# Linguistically Context-Aware Text Corpora Bias Measurement

**Daniel Smarda, EPFL**
daniel.smarda@epfl.ch
*January 15, 2021*
Supervised by:
Navid Rekabsaz (JKU, Austria)
Robert West (DLAB, EPFL, Switzerland)

## Abstract

Identification and removal of bias in word embedding models has gained significant attention in recent years. While several of these methods are creative and effective at removing the chosen definition of bias, most, if not all, of them operate on the assumption that a bias value of 0 is equivalent to the removal of bias. I expand this definition with the assertion that a word is biased if and only if it carries the same connotation as the rest of the words in the language, regardless of if this numerical value is equal to 0. Using the most commonly-used metric of bias and a series of foundational empirical tests, I demonstrate the significance of applying this principal to bias measurement metrics. In particular, I show that the average biases towards common attributes across the English language are generally nonzero to a degree that changes the determination of bias for a target in several social domains. I conduct this analysis for both standalone target sets and the relationship between two target sets and conclude with a discussion of the merits of assessing a social domain from both single-target and double-target perspectives.

## 1 Introduction and Related Work

In the past several years, the study of bias (in particular, gender bias) in natural language systems has ballooned. While several methods have been proposed, the most prominent category of such methods has been cosine-based similarity metrics proposed by Bolukbasi et al. (2016) and Caliskan et al. (2017). Unlike previous studies of large corpora, which were based on word-frequency methods, these authors proposed methods based on the cosine similarity between vectors in word embedding spaces.

Substantial further work in reducing these biases has been proposed (see Bolukbasi et al. (2016) and Zhao et al. (2018)). However, the debiasing research has substantial gaps. Most eloquently, Gonen and Goldberg (2019) demonstrated that even after debiasing, the biases can still be recovered by classifiers, thus demonstrating an only surface-level debiasing of the vectors.

One of the most glaring flaws in the goals of the methods in the above papers is that most debiasing methods focus on reducing bias values to 0. I argue, however, that any definition of language bias must take into account the distribution of language use. Colloquially, this means that a word $w$ is biased if it has the same connotation on average as *the rest of the words in the language*, which is numerically not necessarily equal to 0.

The main contribution of this project is fourfold, and, after a brief theoretical background, this paper is structured accordingly. First, I show that the background distribution of word biases to several attributes follows a mathematically-biased normal distribution. Second, using this conclusion, I assess the bias of individual target sets against attribute words. Third, using a similar method, I calculate the relative bias between two target words and assess its statistical significance as an immediate improvement over the Caliskan et al. (2017) metric. Finally, we show that the difference in bias between two target sets may still be significant when individual bias of target sets shows otherwise and vice versa, indicating the need for more advanced methods of bias measurement.

## 2 Background

Caliskan et al. (2017) clearly defined 10 experiments across several sociological domains. These experiments cover, for example, the pleasantness connotations of names of different races, the permanence connotations of mental and physical diseases, and gender relationships to career/family concepts

and mathematics/science concepts. [1] The choice of terms are based on a body of research from the psychology domain and thus can be considered reasonable representations of the target concepts. In this paper, the words that I use to define the experiments are identical save a few minor exceptions that arose from ambiguities in preprocessing steps in the Caliskan et al. (2017) paper. The titles of the categories used in each of the experiments can be found in Table 1.

### 2.1 Mathematical Definitions

Each experiment is defined by two complementary target concepts $X$ and $Y$ and two complementary attribute concepts $Z$ and $Z'$ against which the target concepts are evaluated. Each concept is defined by a set of word vectors ranging in cardinality from 6 to 25, where $|X| = |Y|$ and $|Z| = |Z'|$.

In each experiment, the bias of a single word is defined as following. Borrowing in notation from Rekabsaz and Hanbury (2018), we have:

$$WEAM(w, Z) = \frac{1}{|Z|} \sum_{x \in Z} \text{cosine}(\boldsymbol{v}_x, \boldsymbol{v}_w) \quad (1)$$

where $WEAM$, the Word Embedding Association Metric, measures the bias of a single word $w$ to a target concept $Z$. The relative bias of a word to the two complementary attribute sets is then given by:

$$\psi(w) = WEAM(w, Z) - WEAM(w, Z') \quad (2)$$

For a multi-word target set $X$, we can define the *standalone bias* of the set as the average of the biases of the terms in the set:

$$\psi_X = \frac{1}{|X|} \sum_{w \in X} \psi(w) \quad (3)$$

and we can calculate the *relative bias* between two target sets $X$ and $Y$ as:

$$\psi_{XY} = \sum_{x \in X} \psi(x) - \sum_{y \in Y} \psi(y) \quad (4)$$

The choice of omitting the division of the set cardinality in the computation of $\psi_{XY}$ was one first published by Caliskan et al. (2017) and maintained in the literature. The choice is preserved here to align with this convention and to avoid the unnecessary division computation.

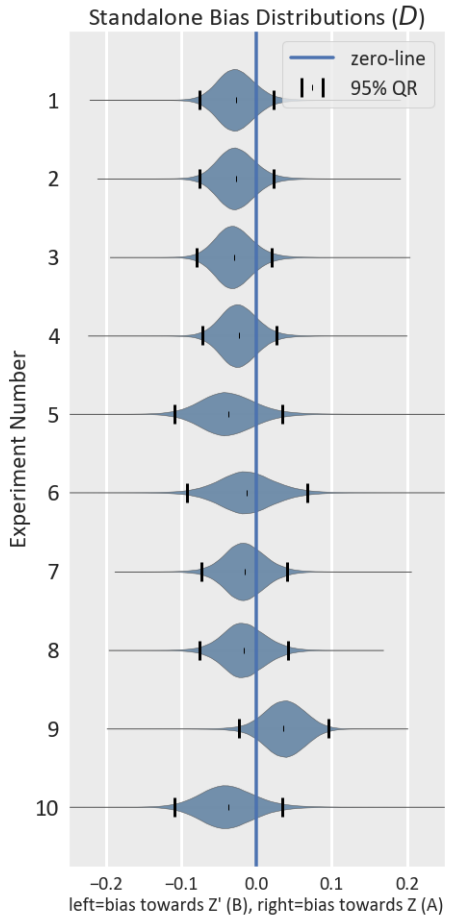


Figure 1: Background Distributions of all words in vocabulary to attribute words for all experiments. The area between the black ticks contains 95% of the distribution density.

## 3 Single-Word Distributions

To assess the nature of the background distribution of the bias of the language towards teh $Z$ and $Z'$ concepts, I first calculated the bias $\psi(w)$ for all words in the language ( 2 million tokens) for each of the 10 experiments. [2]. The distribution of these values for each experiment is shown in Figure 1.

Without even investigating the bias of any particular target words, we can draw several interesting conclusions. First, we notice that the distributions

[1] For the sake of space and focus the justification for the use of the terms in each experiment is omitted here, but the details can be found in the Caliskan et al. (2017) paper.

[2] Due to computational limitations in rendering the diagrams, the distributions only visualize a random sample of 200K words from the language. Computations of values, however, such as those in Table 2, were executed using the entire language

Table 1: **Experiment Definition Labels.** The first two experiments can be considered baseline experiments. The remainder focus on specific problematic stereotypes.

| Experiment # | Target Labels | | Attribute Labels | |
|---|---|---|---|---|
| | $X$ | $Y$ | $Z$ | $Z'$ |
| 1 | Flowers | Insects | Pleasant | Unpleasant |
| 2 | Instruments | Weapons | Pleasant | Unpleasant |
| 3 | European-American Names | African-American Names | Pleasant | Unpleasant |
| 4 | European-American Names | African American Names | Pleasant | Unpleasant |
| 5 | European-American Names | African American Names | Pleasant | Unpleasant |
| 6 | Male Names | Female Names | Career | Family |
| 7 | Math | Arts | Male Terms | Female Terms |
| 8 | Science | Arts | Male Terms | Female Terms |
| 9 | Mental Disease | Physical Disease | Temporary | Permanent |
| 10 | Young People's Names | Old People's Names | Pleasant | Unpleasant |

appear to be roughly normally shaped. Given the "natural" origin of language and its use, this is not an unsurprising conclusion. However, it is notable that the distribution is approximately symmetric and not, for example, skewed as one might hypothesis given the tone-homogeneous nature of single-source corpora.

We also notice that the average $\psi$ value is clearly nonzero in every experiment. In the figure, a mean that is to the left of the origin indicates that all words in the vocabulary (i.e., language) are closer in the vector space to the $Z'$ concept than the $Z$ concept. In Experiment #4, for example, where $Z'$=`unpleasant` and $Z$=`pleasant`, this means that the word vectors are closer in direction to the `unpleasant` vector than the `unpleasant` vector. We notice that every distribution is translated from the origin. The direction of bias in the experiments only carries meaning insofar as the original target sets $X$ and $Y$ were compared (which was not considered in the development of this plot). Nonetheless, the existence of bias of qualitatively at least a standard deviation in each experiment shows that off-center bias is consistent, and not restricted only to a few attributes.

## 4 Bias of Standalone Target Sets

### 4.1 $\Phi$ Values

To analyze the effect of the the nonzero means of the above distributions, I fit a normal distribution with the parameters obtained from the distributions in section 3 and calculate the CDF $\Phi$ for the target sets defined in Table 1. Because the distribution is bi-directional, a very small $\Phi$ value indicates a very strong bias towards the concept $Z'$ and a very large $\Phi$ value indicates a very strong bias towards the concept $Z$.

Table 2 compares the $\Phi$ values of the target sets of each of the experiments calculated from the $\mu = 0$ assumption to those calculated from the distributions in section 3, where $\mu = \psi_X$. Results vary by experiment (i.e., no homogeneous trend is clear). But, we see significant effects in several specific instances. For example, using the definition that a word is biased if it is in the top or bottom 2.5% of words in either direction (akin to $\alpha = 0.5$), we find that:

- The X-term `Science` in Experiment 8 is unbiased ($\Phi = 0.91$) if we assume $\mu = 0$ but nearly biased toward the $Z'$ term `Male` if we assume $\mu = \psi_X$ ($\Phi = 0.971$).
- The Y-term `Old People's Names` in Experiment 10 is unbiased ($\Phi = 0.90$) if we assume $\mu = 0$ but biased towards (perhaps surprisingly) the $Z$ term `Unpleasant` if we assume $\mu = \psi_X$ ($\Phi = 0.99$).
- The X-term `Mental Disease` in Experiment 9 is arguably unbiased if we assume $\mu = 0$ ($\Phi = 0.03$) but very clearly biased towards the $Z'$ term `Unpleasant` if we assume $\mu = \psi_X$ ($\Phi = 0.001$).

### 4.2 Visualizations

The results of subsection 4.1 are possibly more easily digested in visual form. To accommodate this, I plotted the bias of each individual word in the experiment against the background distribution of biases (the distributions in Figure 1). While the plots contain a level of detail disproportionate to the main text of this paper and are thus relegated

Table 2: Φ Value Comparisons

| Experiment # | X-Terms | | Y-Terms | |
|---|---|---|---|---|
| | $\mu$=$\psi_X$ | $\mu$=0 | $\mu$=$\psi_X$ | $\mu$=0 |
| 1 | 0.996 | 0.949 | 0.227 | 0.041 |
| 2 | 0.998 | 0.973 | 0.306 | 0.067 |
| 3 | 0.993 | 0.913 | 0.641 | 0.231 |
| 4 | 0.962 | 0.811 | 0.590 | 0.255 |
| 5 | 0.999 | 0.989 | 0.948 | 0.744 |
| 6 | 0.794 | 0.694 | 0.001 | $10^{-4}$ |
| 7 | 0.823 | 0.640 | 0.526 | 0.307 |
| 8 | 0.971 | 0.910 | 0.684 | 0.468 |
| 9 | 0.001 | 0.033 | $10^{-10}$ | $10^{-7}$ |
| 10 | $10^{-4}$ | 0.994 | 0.987 | 0.899 |

to Appendix D, for comprehensive understanding I highly recommended viewing them alongside Table 2.

## 5 Relative Bias of Two Target Sets

The Φ values from section 4 and the visualizations in Appendix D indicated that in some experiments (e.g., Experiment #4), the means of both target sets $X$ and $Y$ were unbiased, and in others (e.g., #9), both target sets were biased in the same direction. In both of these cases, one might expect that the relative bias (see Equation 4) would also exhibit unbiased behavior. However, I experimentally show that this is not the case.

We want to test the null hypothesis that the relative bias between the target sets $X$ and $Y$ is the average difference between two sets of words in the population. Caliskan et al. (2017), who first proposed a test of a similar goal, generated the distribution of the test statistic by conducting a permutation test. However, the researchers only calculated the distribution over all equal-size partitions of $X \cup Y$. I offer an alternative definition that takes into consideration the background distribution of relative biases in the language. Note that this can be considered an empirical parallel work to Ethayarajh et al. (2019), who demonstrated the inclination of the Caliskan et al. (2017) test to overestimate bias.

Consider, from Equation 4, the bias between two arbitrary sets of words $\Xi$ and $\Upsilon$ as $\phi_{\Xi\Upsilon}$. Since $\Xi$ and $\Upsilon$ are sets of $|\Xi| = |\Upsilon| = n$ terms, we can generate up to $\binom{|V|}{n}$ similar target sets from our vocabulary $V$. If this size-$\binom{|V|}{n}$ group of all possible target sets is denoted $A$, we can then say that the distribution of relative biases across the languages to the targets $Z$ and $Z'$ is comprised of $\{\phi_{\Xi\Upsilon} : \Xi \in A, \Upsilon \in A\}$, or, in equivalent terms, $\{\phi_{\Xi\Upsilon} : (\Xi, \Upsilon) \in A \times A\}$.

For each of the experiments, I randomly sampled and calculated $\psi_{\Xi_i\Upsilon_i}$ for $i \in [1, 100\text{K}]$ and plot the results in Figure 2. The values displayed on the figure are analogous to the Φ values in the $\mu$=$\psi_X$ columns of Table 2, and represent the proportion of word pairs with a numerically-lesser relative bias than the value $\psi_{XY}$, using the $X$ and $Y$ sets from Table 1. In an isolated manner, given other research in this field, the numerical results are not altogether surprising, as all except perhaps arguably one of the experiments show significant bias, which is roughly the same conclusion reached by Caliskan et al. (2017) (for numerical comparisons to the Caliskan et al. (2017) method, see Appendix C). The most interesting conclusions come when we combine the results of the standalone bias experiments and the relative bias experiments, as discussed in the next section.

## 6 Cumulative Results

When we combine the results from sections 4 and 5, we notice two particular cases which indicate the need for more detailed bias analysis.

**Unbiased Individual Sets with Significant Relative Bias** . Consider Experiment #7, the detailed plots for which can been seen in Figure 3. If we look at the figure, or the Φ values in Table 2, we notice that neither target set `European-American Names` nor `African-American Names` show a bias towards the attribute pair `Pleasant/Unpleasant`. However, when

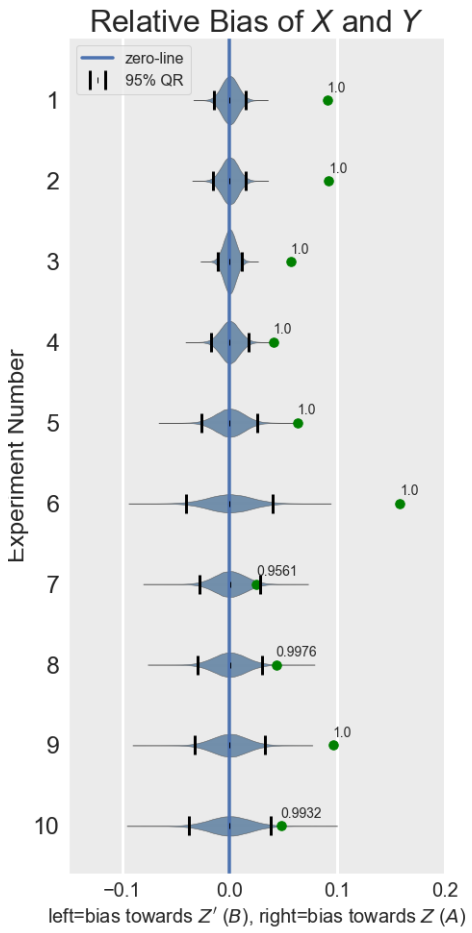


Figure 2: Distribution of relative bias between random word sets, and $\Phi$ value for the experiment sets.

analyzing the difference in biases between these sets (Figure 2), we see that the relative bias is clearly significant. This trend is also observable in Experiment #8 and (though the relative bias is less-significant) Experiment #7.

**Target Sets Biased in the Same Direction** Consider Experiment #9 (Figure 4). Both target sets `Mental Disease` and `Physical Diseases` exhibit a bias in the same direction: towards `Pleasant`. At first observation, this may not seem problematic, but again the relative bias is significant as described in section 5. This same trend is noticeable in Experiment #10.

**Identification of Individually Problematic Target Sets** Finally, the combination of standalone and relative biases is an important addition to the current literature because, from the perspective of sociology, it identifies both the existence of relative bias, and the root cause for such bias. Most texts on social justice (consider, among many, many examples, Johnson (2018) or Adams et al. (2013)) require a preference of one group over another as a prerequisite for bias to exist (explored in section 5). Alone, however, this information fails to diagnose the problem. As a final case, consider Experiment #3. The individual experiments conducted as described in section 4 demonstrate that in fact `African-American Names` are treated, on average, no differently than other words in the language, and that the solution to correcting for this racial bias in in fact removing the excessive `Pleasant` connotation of `European-American Names` rather than making `African-American Names` more `Pleasant`.

## 7 Conclusions and Further Work

Most work on debiasing word embeddings based on cosine similarity-based metrics focus on reducing the bias metric to 0. However, I have shown that the background distribution of biases of words in the language are not centered around zero. Furthermore, I have demonstrated that the effects of these shifts on statistical bias determinations are not insignificant. As bias has been significantly demonstrated and several multi-faceted tests of bias proposed in this paper and others, the most pressing next steps are those focused on sociological impact. This may take the form of additional experiments, or more up-to-date experiments. Or, it may take the form of philosophical discussion and incorporation of research from linguistics and social justice to solidify the most applicable definitions of word embedding corpora bias to practical applications.

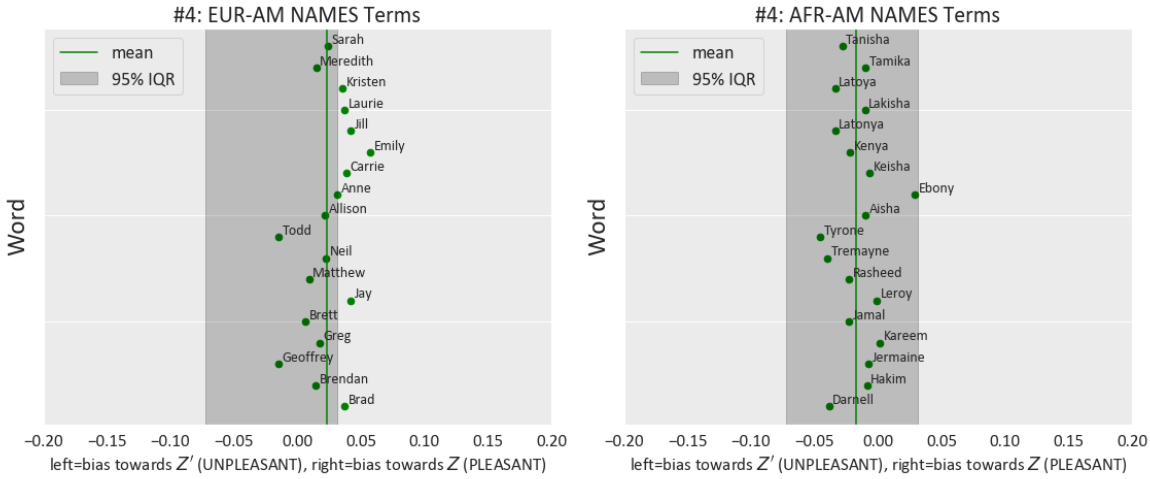


Figure 3: **Experiment #4** The grey regions correspond to the regions indicated by the black ticks in Figure 1. Words (and word sets) that fall within this region can be considered to be unbiased with respect to the rest of the words in the language.

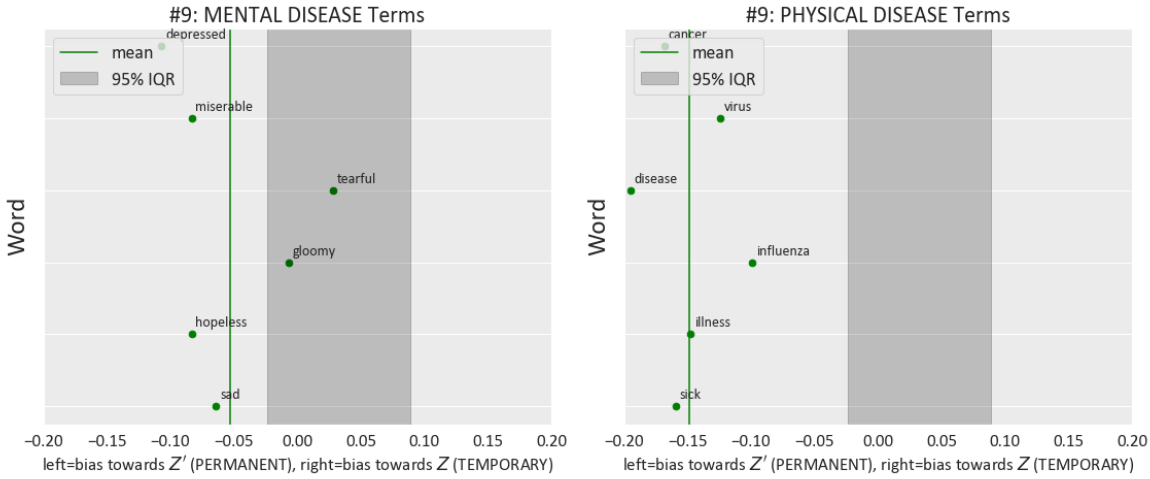


Figure 4: Experiment #9

## References


Maurianne. Adams, Warren J. Blumenfeld, Carmelita Rosie Castaneda, Heather W. Hackman, Madeline L. Peters, and Ximena. Zuniga. 2013. *Readings for diversity and social justice*. Routledge, New York.

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 *Proceedings of the 30th International Conference on Neural Information Processing Systems*.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*.

Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. 2019. Understanding undesirable word embedding associations. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *Proceedings of the North American Chapter of the Association for Computational Linguistics*.

Inc. Google. 2013. *word2vec*.

Allan G. Johnson. 2018. *Privilege, Power, and Difference*, third edition. edition. McGraw-Hill Education, New York, NY. HOLLIS number: 990151371260203941.

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.

Navid Rekabsaz and Allan Hanbury. 2018. An unbiased approach to quantification of gender inclination using interpretable word representations. *CoRR*, abs/1812.10424.

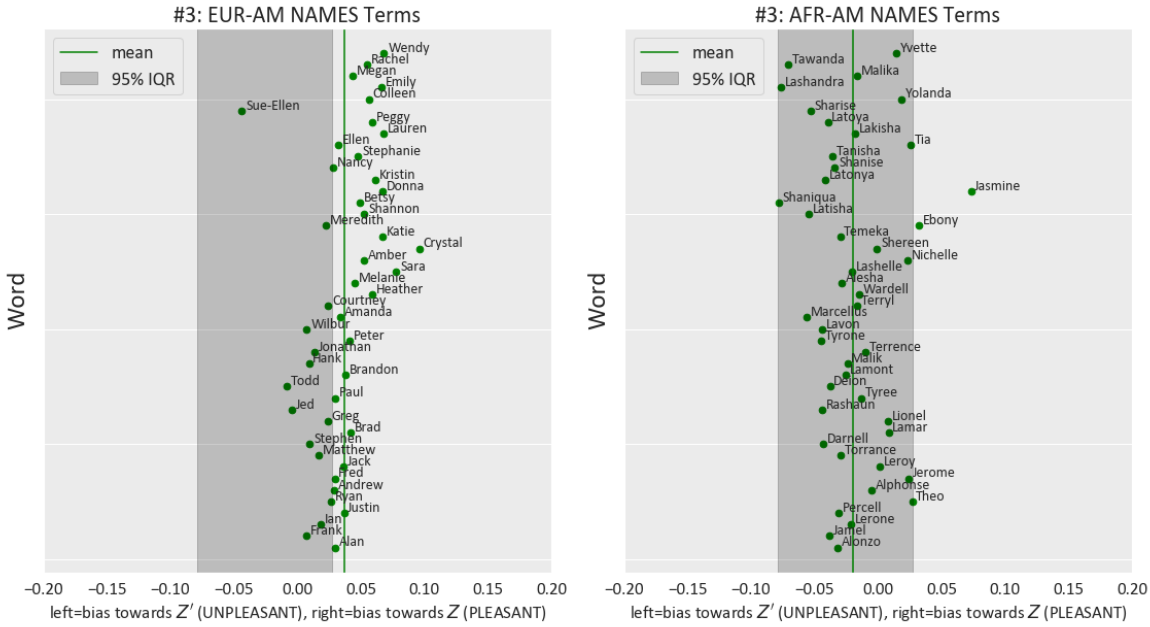


Figure 5: Experiment #3

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018. Learning gender-neutral word embeddings.

## A Source Code

The computations for this project were implemented using Python (NumPy, SciPy, statsmodels, gensim, and Seaborn). The code is in a repository that is currently private but can be made available upon request to `daniel.smarda@epfl.ch`.

## B Datasets

All of the numerical results listed in this paper correspond to experiments run on the 2.2-million-token Common Crawl word embedding model from the GloVe project (Pennington et al., 2014). This is the same model presented in the main paper by Caliskan et al. (2017). Importantly, we also ran these tests on the common Google News word2vec model (Google, 2013) with very similar results.

## C Relative Bias Values

The table below compares the relative bias $\Phi$ values calculated in section 5 to the $p$-values obtained by Caliskan et al. (2017). From a high level, we reach the same conclusions as Caliskan et al. (2017): that the relative bias for all experiment choices (except for, arguably, #7) is significant.

Table 3: Comparison of our $\Phi$ and Caliskan et al. (2017) $p$

| Experiment # | $\Phi$ (this paper) | $p$ (Caliskan) |
|---|---|---|
| 1 | $< 10^{-10}$ | $10^{-7}$ |
| 2 | $< 10^{-10}$ | $10^{-7}$ |
| 3 | $< 10^{-10}$ | $10^{-8}$ |
| 4 | $10^{-6}$ | $10^{-4}$ |
| 5 | $10^{-6}$ | $10^{-3}$ |
| 6 | $< 10^{-10}$ | $10^{-3}$ |
| 7 | .044 | .018 |
| 8 | .002 | $10^{-2}$ |
| 9 | $10^{-8}$ | $10^{-2}$ |
| 10 | .007 | $10^{-2}$ |

## D Multi-word Plots

This appendix contains the detailed word bias plots from section 4 for all 10 experiments. The grey intervals are the same intervals demarcated by the ticks in Figure 1.

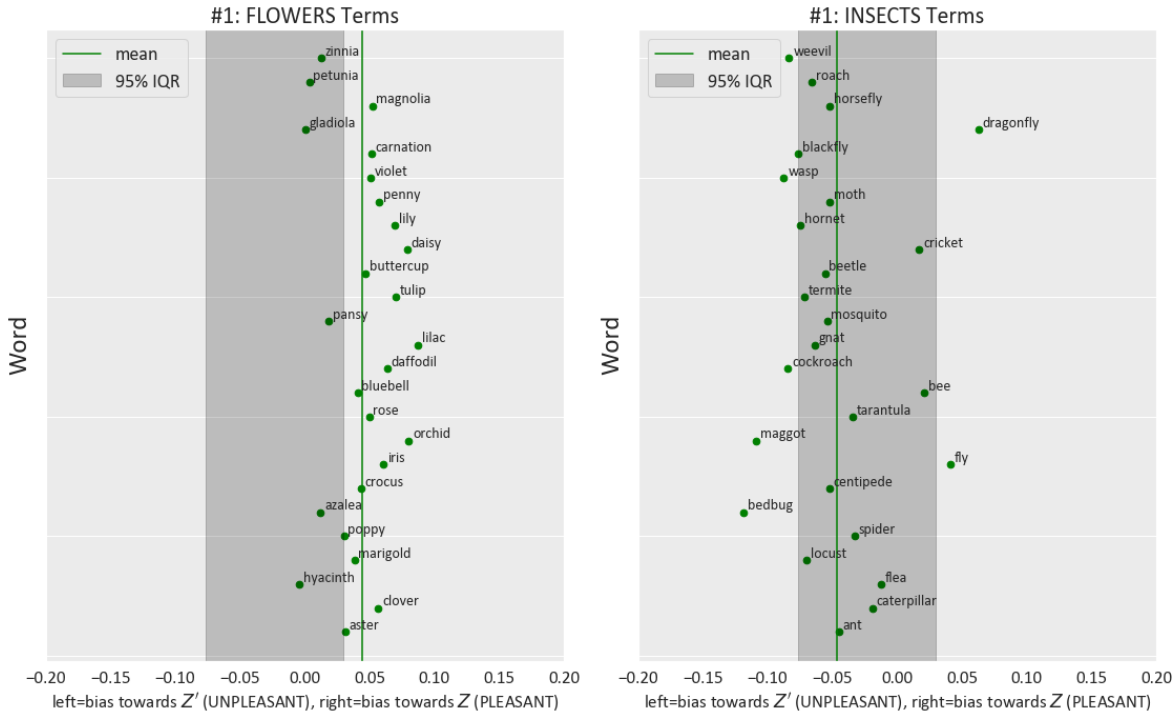


Figure 6: Experiment #1

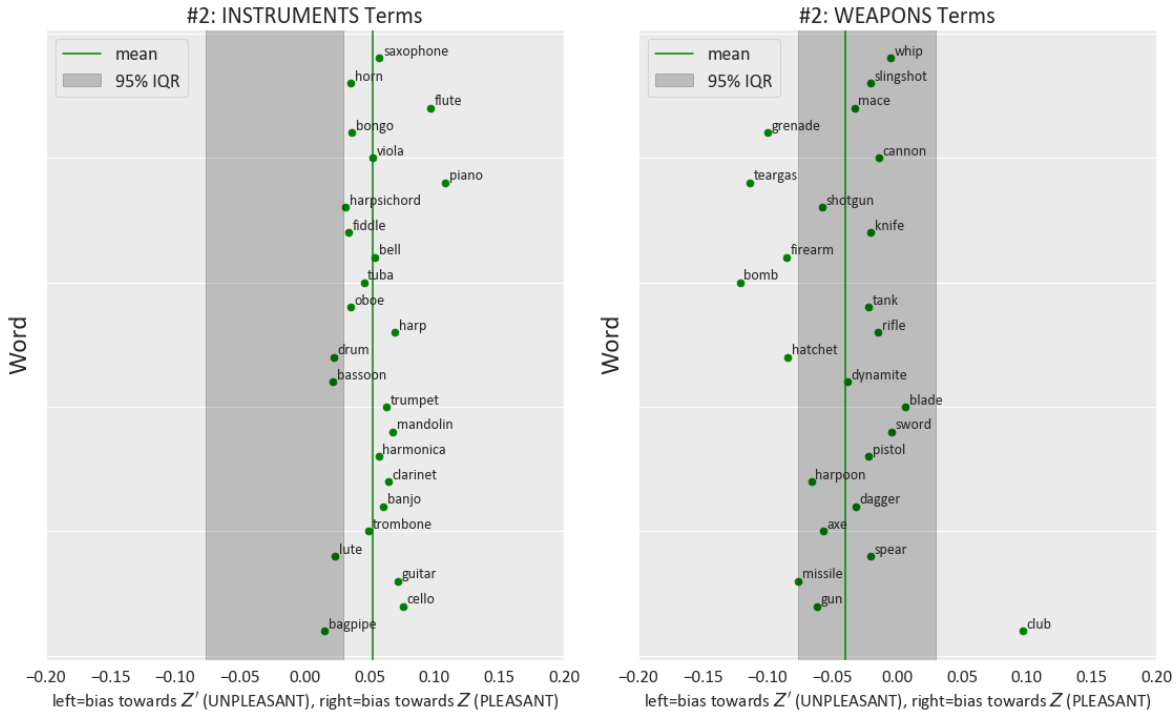


Figure 7: Experiment #2

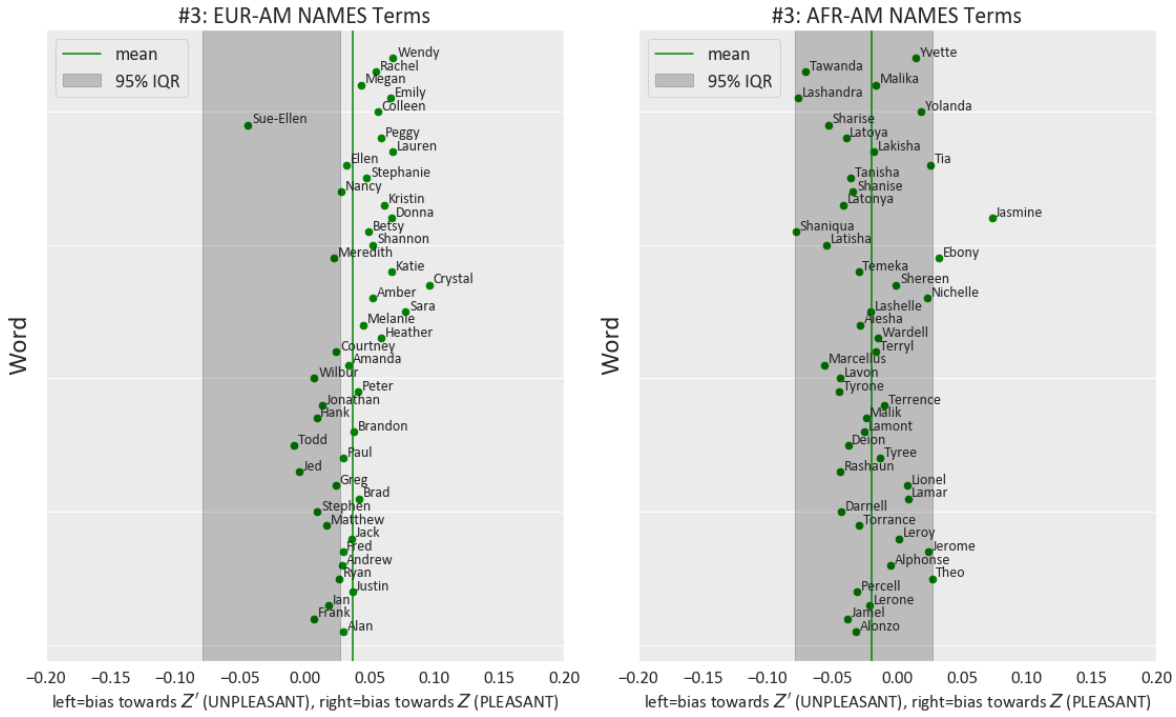


Figure 8: Experiment #3

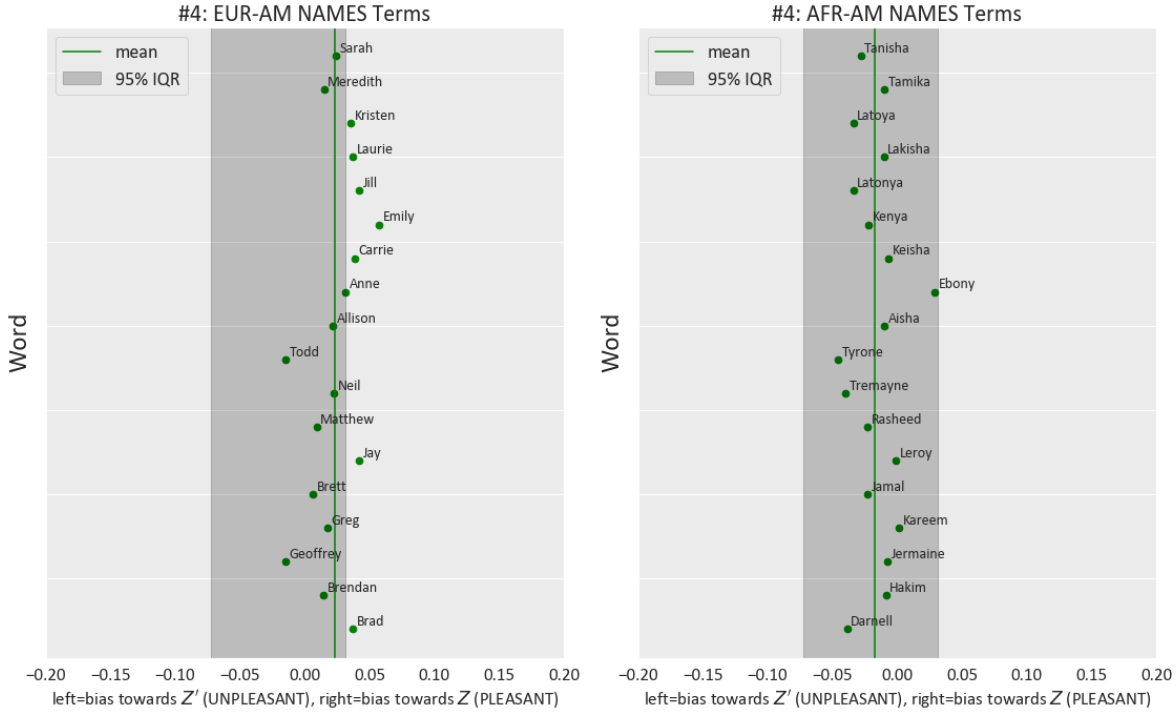


Figure 9: Experiment #4

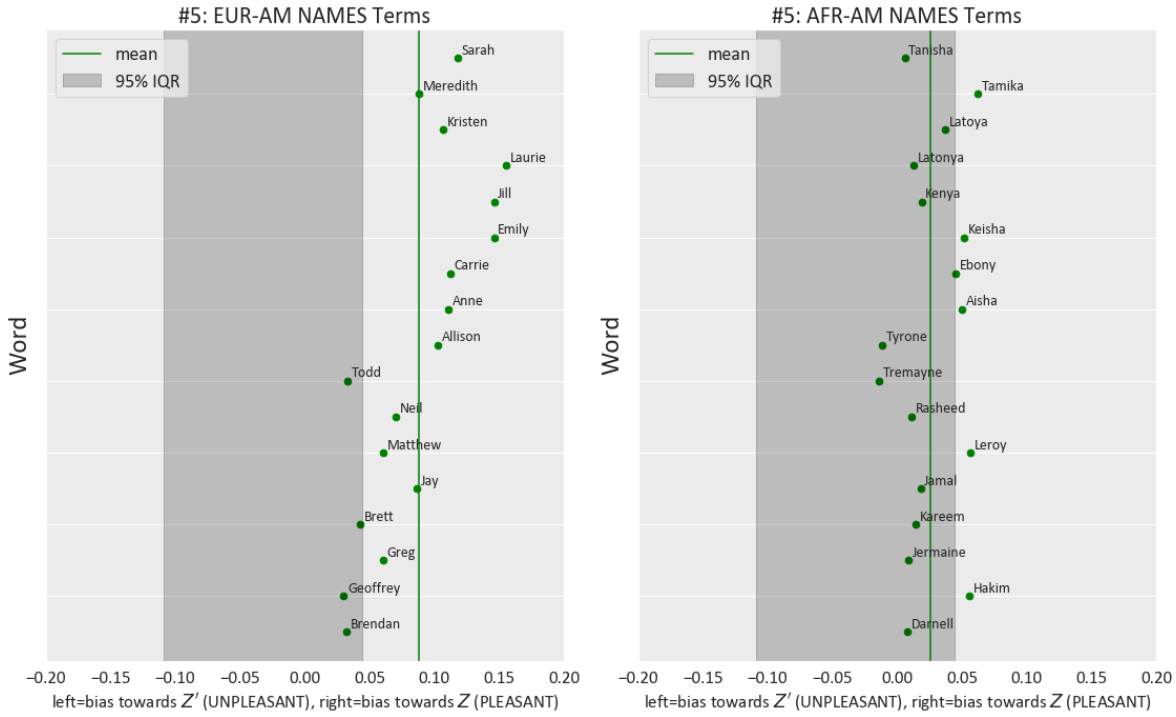


Figure 10: Experiment #5

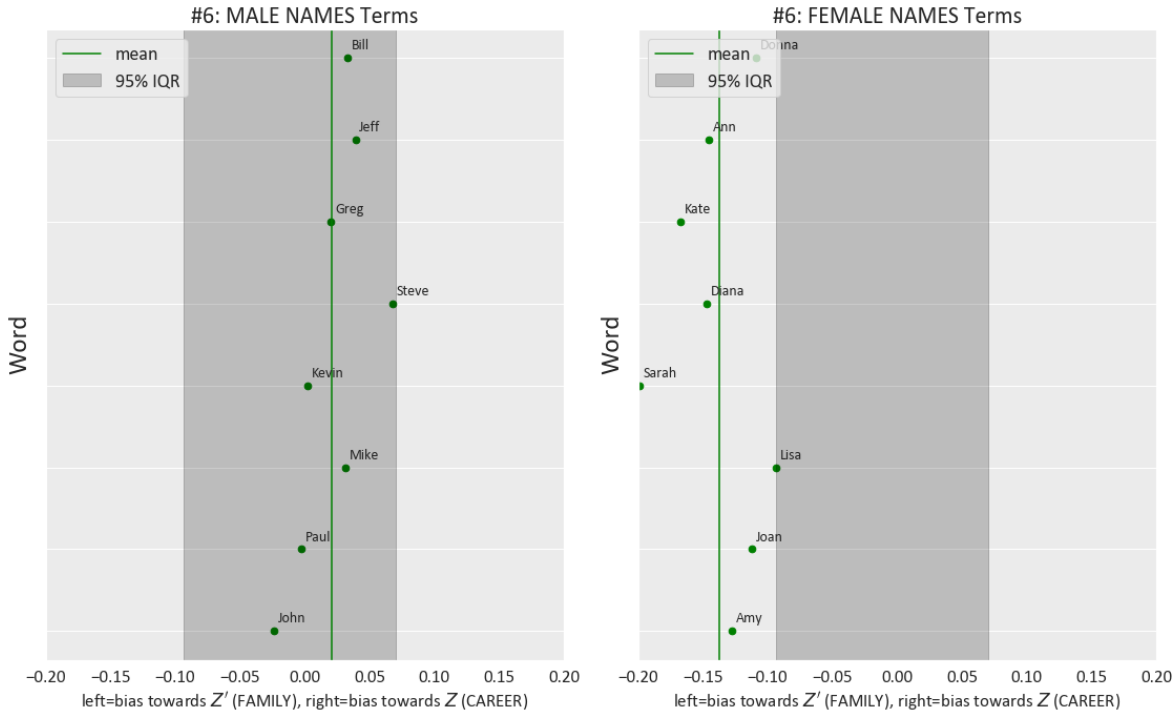


Figure 11: Experiment #6

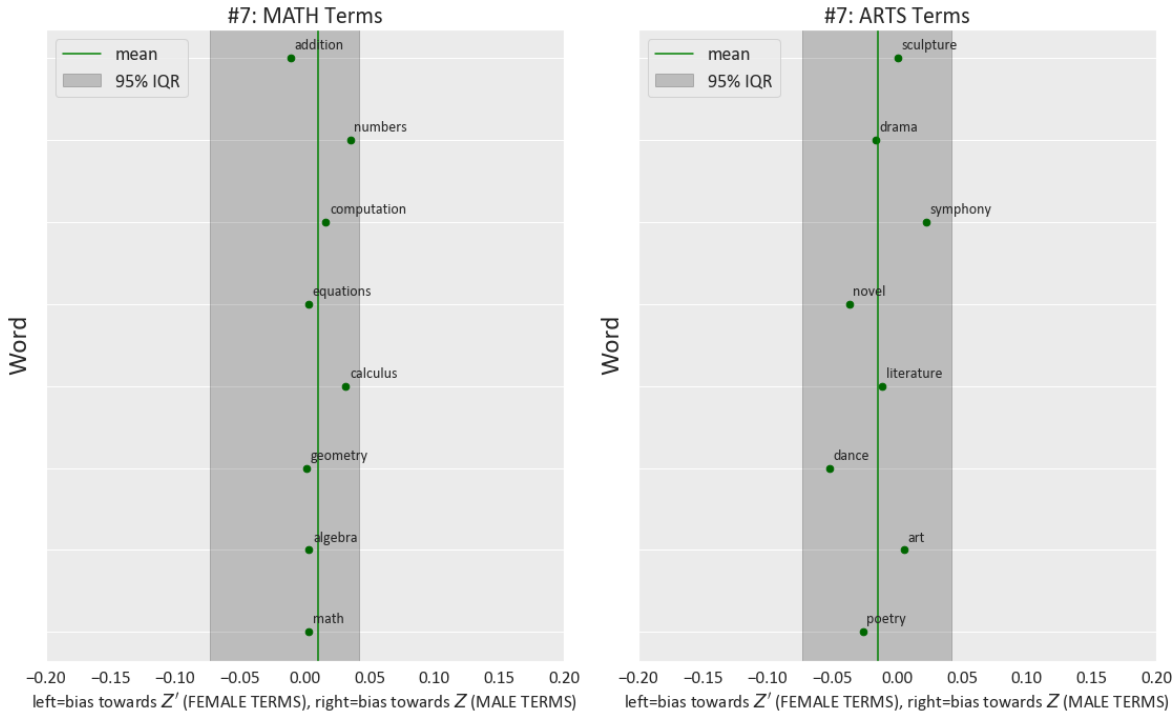


Figure 12: Experiment #7

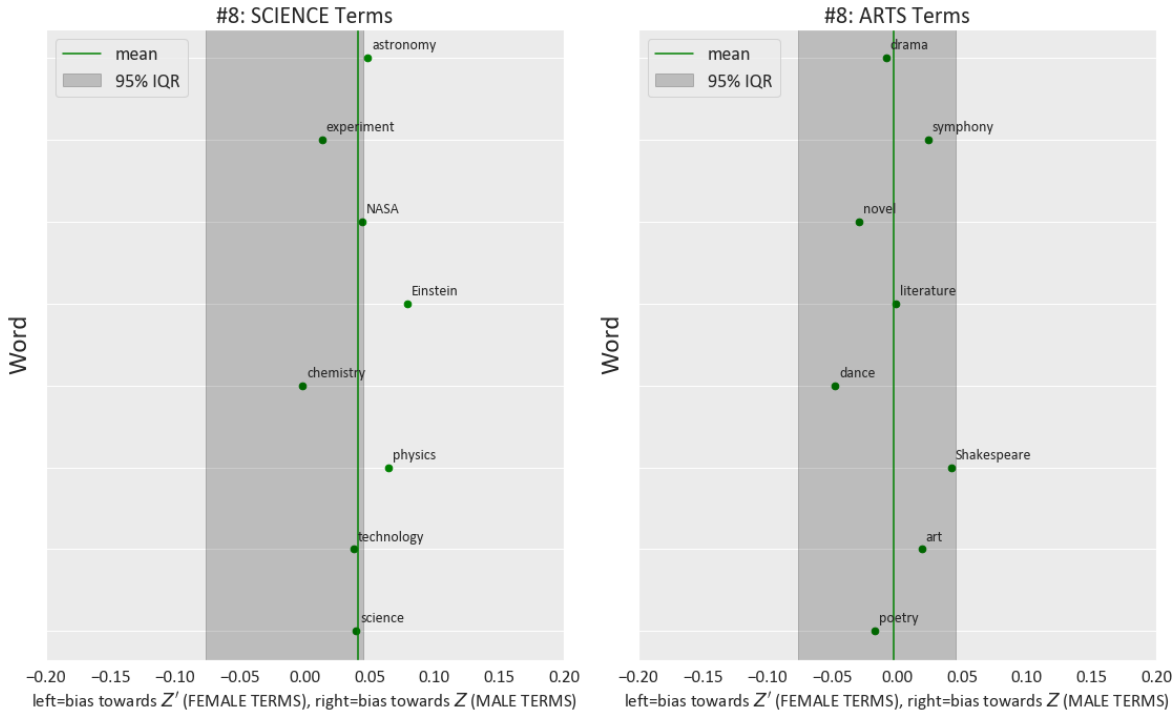


Figure 13: Experiment #8

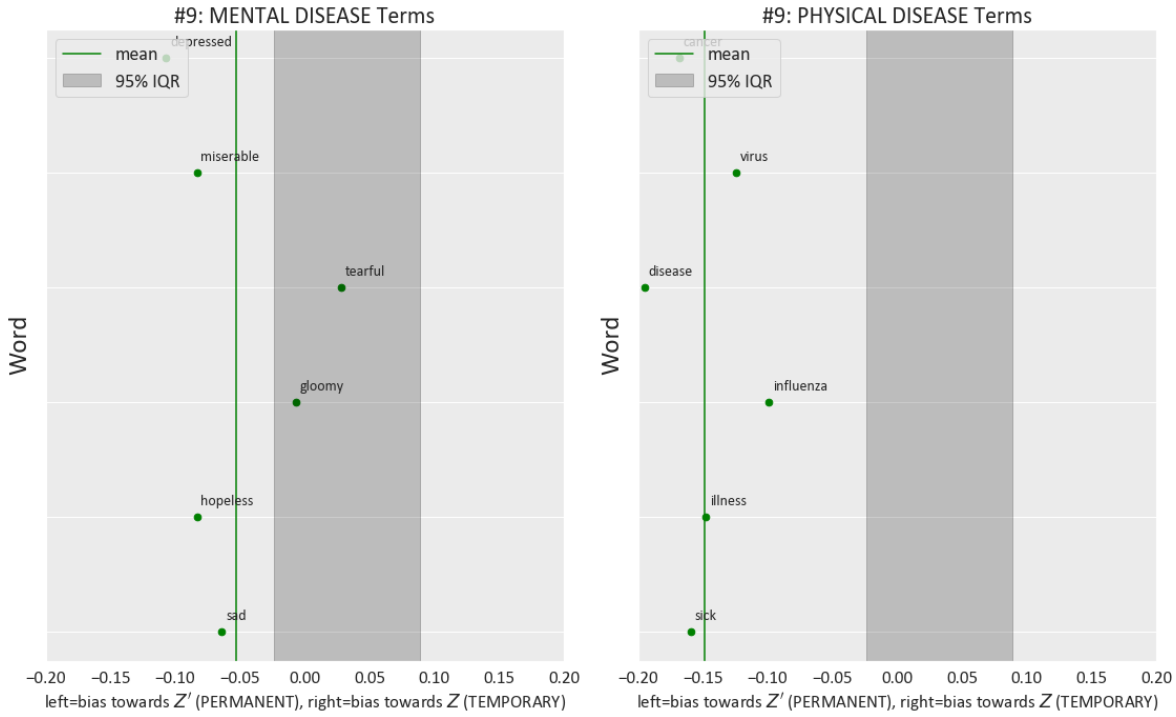


Figure 14: Experiment #9

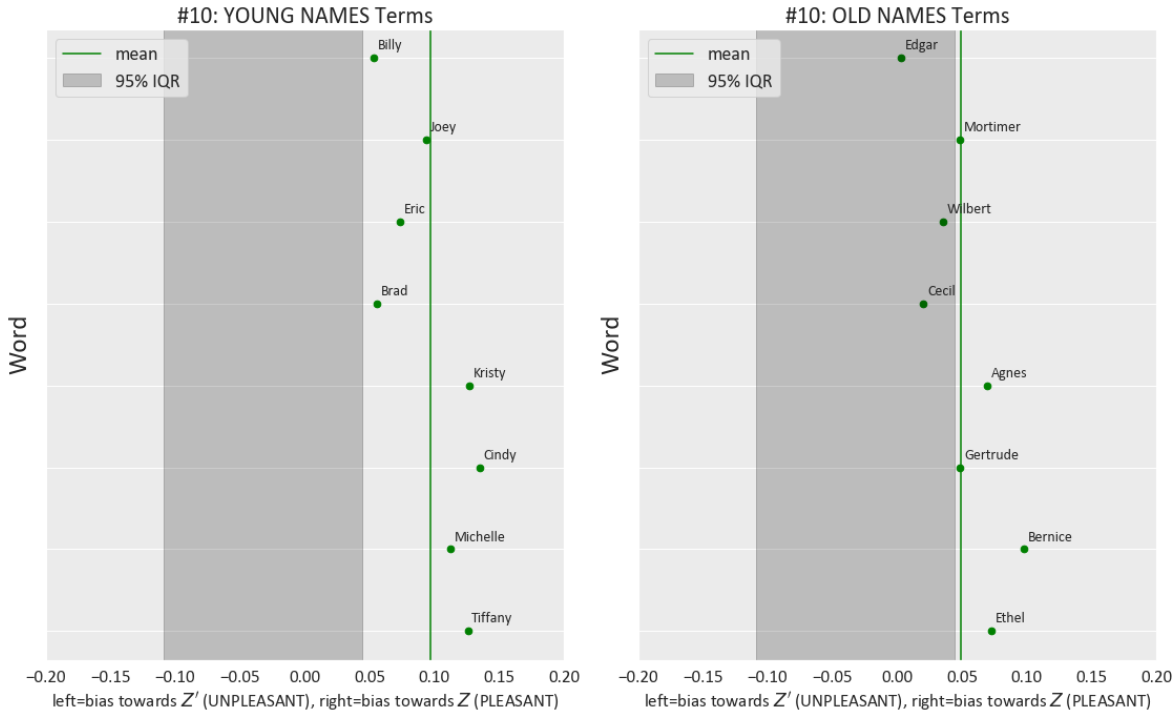


Figure 15: Experiment #10