sociological science

Supplement to:

Jones, Jason J., Mohammad Ruhul Amin, Jessica Kim, and Steven Skiena. 2019. “Stereotypical Gender Associations in Language Have Decreased Over Time.” Sociological Science 7: 1-35.

APPENDIX. Further details concerning the word embeddings method.

The HistWords Embeddings

HistWords is a collection of tools and historical word embeddings datasets (Hamilton, Leskovec & Jurafsky, 2016). The tools were originally created to quantify changes in word meanings over time. In this manuscript, we inspected word embeddings trained separately on text from books published in English from each decade in the range 1800-2000. The Google Books NGram Corpus provided the text.

The embeddings were constructed using the skip-gram with negative sampling method known as word2vec (Mikolov et.al. 2013). Minimal pre-processing was applied – specifically, each word in the text was lowercased and stripped of punctuation. Words that occurred fewer than 500 times in the decade's text were discarded. No lemmatization or stemming was applied.

Word vectors of 300 dimensions resulted from applying the word2vec algorithm to each decade of text. Hyperparameters were set following the recommendations of Levy, Goldberg & Dagan (2015).

As the word embeddings for all decades were generated independently of each other, the HistWords process includes a further step to align the set of embeddings into the same multidimensional coordinates. This is necessary because the stochastic nature of word vector optimization will generate an embedding space for each decade, and the spaces are not naturally aligned. In particular, the HistWords process used orthogonal Procrustes to align the learned decade-level embeddings to the word embeddings generated from the final decade (the 1990s). This unified word embedding space can be analyzed to compare relative displacements of word vectors across different time periods. In this paper, we used these aligned historical word embeddings to compute the gender-to-domain associations over time.

Word-to-Word Distance and Similarity

This section will further discuss the calculation of word associations within word embeddings. The code used for calculations is available at https://github.com/ruhulsbu/StereotypicalGenderAssociationsInLanguage.

In a word embedding, every word is individually represented by a vector in one multidimensional space. The natural metric to measure distances in such a space is the cosine distance. One way to think of the value of the cosine distance is as the rotation necessary to bring two vectors into alignment. If two vectors have a cosine distance of 0, they point in exactly the same direction in the embedding space. The greater the difference in the vectors' alignment, the larger the cosine distance. For any two words, say "man" and "office," there exists a cosine distance between their corresponding word vectors in the embedding.

Similarity is the complement of distance – i.e. similarity = (1 – |distance|). When distance is at minimum, similarity is at maximum. The similarity of two words (what we call the association between word1 and word2) is thus 1 – |cosine_distance(word1, word2)|.

Gender-to-Domain Similarity

Gender-to-domain similarity is the average of the cosine similarity measured between all possible pairs of gender words and domain words. This measure can be defined mathematically as follows:

$$gender\_to\_domain\_similarity = \frac{1}{n}\sum_{1\leq i\leq n}(\frac{1}{m}\sum_{1\leq j\leq m} cosine\_similarity(G_i, D_j))$$

Here, G_i = embedding of i^{th} word from gender G, where total number of gender words is n.
Here, D_j = embedding of j^{th} word from domain D, where total number of domain words is m

In this paper, we have derived our analysis from the gender to domain similarity of two genders (i.e. male and female) and four domains (i.e. family, career, science and arts). Thus, the gender to domain similarity between “female” and “family” essentially represents how closely these two concepts are associated. The higher the gender to domain similarity, the higher the association between two concepts.

Male Gender Bias

Male Gender Bias is a measure of how much more the male gender is associated to a given domain than the female gender to that domain.

$$male\_bias_D = male\_to\_domain\_similarity_D - female\_to\_domain\_similarity_D$$

Thus for a domain *D*, male gender bias can be measured by subtracting gender to domain similarity of female from that of male. For example, for the domain “family”, we measure the male gender bias by subtracting the association between “female” and “family” from the association between “male” and “family”. If the male gender bias is positive then it tells us that the domain “family” is more associated or biased towards “male” than “female”. On the other hand, if the male gender bias is negative then it means the opposite; the domain “family” is biased towards “female” rather than “male”.

Contextualizing the Magnitude of Observed Changes in Association

This section will discuss the relative magnitude of changes in association observed over time. We performed the following simulation to compare the results for the gender-domain pairs discussed in this manuscript to the expected variation across the decade-by-decade word embeddings from Hamilton, Leskovec & Jurafsky, 2016. To derive a baseline level of variation over time, we constructed “concepts” by creating word lists consisting of randomly chosen words. Then for a pair of concepts, we measured the change in association similarity (see male gender bias above) across the decades. This resulted in a plot, showing the bias towards a concept over time, for which we obtained a regression line as well as the slope. We repeated this process in 1000 independent simulations. The histogram in Figure A1 summarizes the distribution of slopes for random concepts. For comparison, we show four vertical lines representing the gender-domain associations discussed in this manuscript. The changes over time for these gender-domain associations are outliers. This implies that the changes we observe are not of the magnitude expected by chance when comparing word embeddings over time.

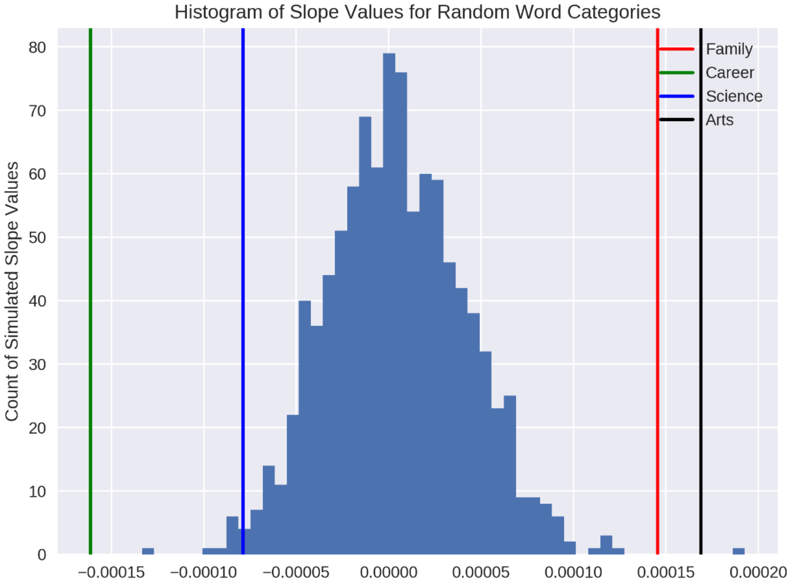

Figure A1. The histogram represents changes in associations over time when concepts are constructed using lists of words chosen at random. The four vertical lines represent the changes in association discussed in this manuscript.

Temporal Granularity

We did not compute Male Gender Bias on a yearly basis for two reasons. First, aggregating to decades decreases noise in the measurement. Calculating word embeddings requires large amounts of text. The less text input, the noisier the word embedding language model will be. Aggregating to decades allows for inspection of trends over two centuries at a reasonable granularity without introducing the noisiness of single-year effects. Second, aggregating to decades simplifies the analysis and allows for better replicability. Training and aligning word embeddings for every year would be very computationally expensive. The decade-level embeddings had already been calculated and are available to download for those seeking to repeat our analysis or conduct their own.

Maintaining Comparable Measurement over Time

More than one reviewer has made the observation that words change in meaning over time and suggested that our analysis should thus be made over word lists that also change over time. We disagree for several reasons. First, keeping the word lists static simplifies interpretation. If the word lists changed over the decades at the same time the embeddings were changing, it would be unclear how to interpret any change in the computed association scores. Did the score change because the association between concepts changed or because the concepts are now represented differently with different words?

Second, to the authors' knowledge there exists no principled, systematic method to select the "correct" words at each particular decade. Which words should denote Science in 1830 versus 1930? We might imagine our own lists, but they would likely differ from those imagined by others. We might use the embeddings to find clusters of words at each decade, but using the language model to both define the changing meaning of a domain and simultaneously its changing association with other domains introduces the confound discussed previously.

Third, these word lists allow comparability to previous research. These word lists were originally developed to test associations in human minds. Because millions of tests have been performed with these word lists, we can compare cross-sectional differences in human results with the trends observed in this work. Additionally, the previous work on gender bias within word embeddings (e.g. Caliskan, Bryson, and Narayanan 2017) used these word lists, and the present work adds context to these results best if the same word lists are used.

Removal of Proper Nouns from Word Lists

Reviewers have mentioned the peculiarity of proper nouns in the Arts and Science domain word lists. These terms were present in the lists used in previous research, and we therefore include them in our primary analysis to maintain comparability with previous work. Below, however, we repeat our analysis after removing the term “Shakespeare” from the Arts list and removing “NASA” and “Einstein” from the Science list. The results are substantively equivalent to those previously presented.

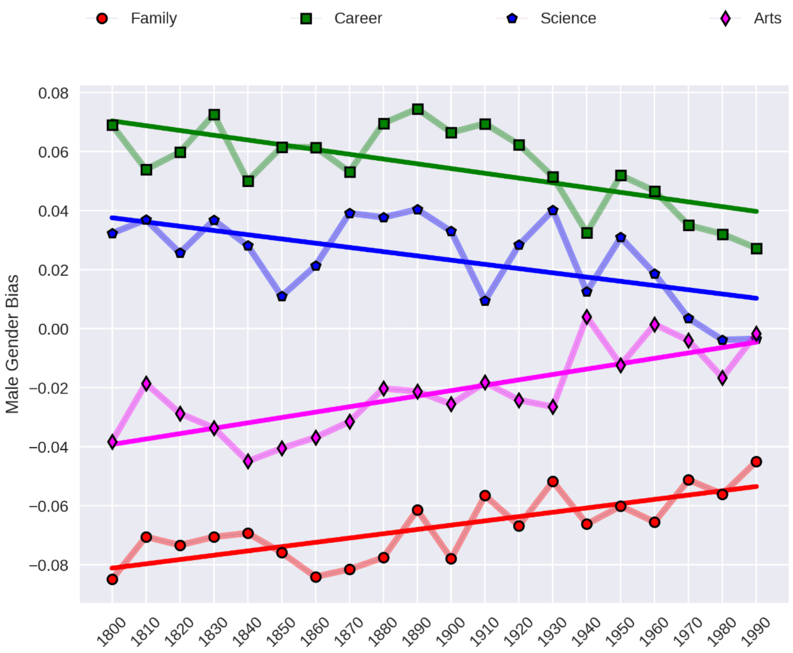

Figure A2. After proper nouns were removed from the Arts and Science lists, Male Gender Bias in each decade of text for the four domains of Family, Career, Science and Arts. Male Gender Bias is calculated by subtracting Female cosine similarity from the Male cosine similarity per each domain. Compare to Figure 2 in the main text. The x-axis is labelled with the first year of each decade – e.g. the label "1910" is used to mark the measurement based on the word embedding trained on text from 1910 up to but not including 1920.

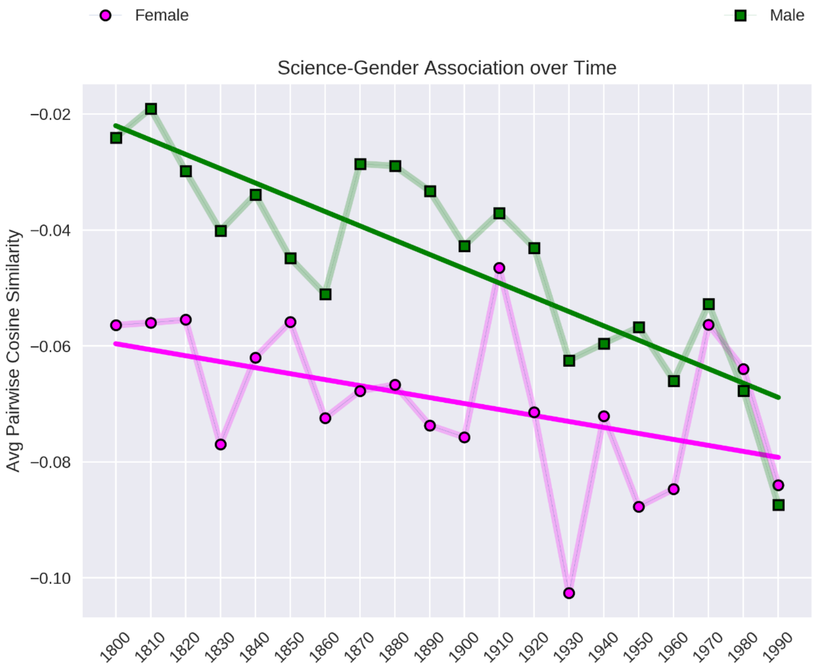

Figure A3. After the proper nouns NASA and Einstein were removed from the Science word list, Female association and Male association with the domain Science measured by decade. Compare to Figure 5 in the main text. The x-axis is labelled with the first year of each decade – e.g. the label "1910" is used to mark the measurement based on the word embedding trained on text from 1910 up to but not including 1920.

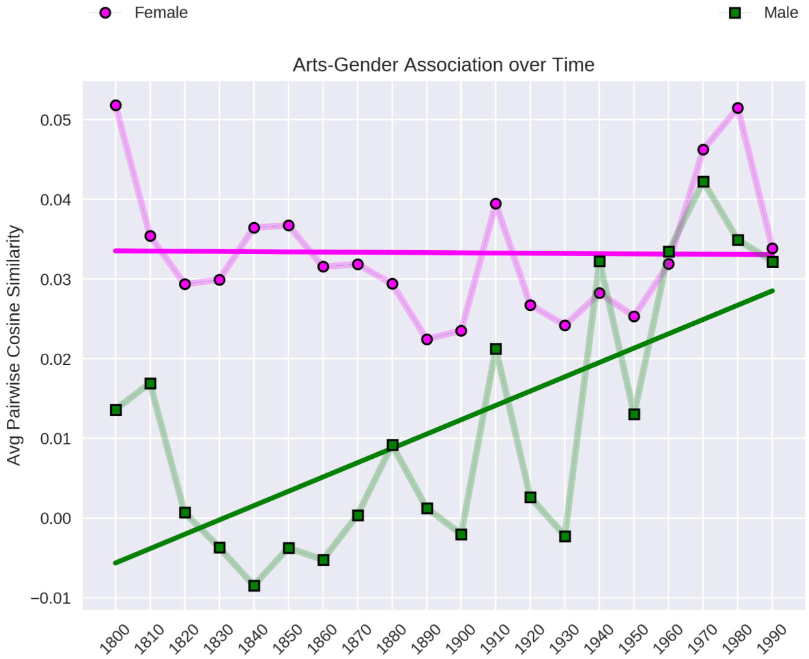

Figure A4. After the proper noun Shakespeare was removed from the Arts word list, Female association and Male association with the domain Arts measured by decade. Compare to Figure 6 in the main text. The x-axis is labelled with the first year of each decade – e.g. the label "1910" is used to mark the measurement based on the word embedding trained on text from 1910 up to but not including 1920.

References

Caliskan, Aylin, Joanna J. Bryson, and Arvind Narayanan. 2017. “Semantics Derived Automatically from Language Corpora Contain Human-Like Biases.” *Science* 356:183–6.

Hamilton, William L., Jure Leskovec, and Dan Jurafsky. 2016. “Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change.” Pp. 1489–501 in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Stroudsburg, PA: Associatoin for Computational Linguistics.

Levy, Omer, Yoav Goldberg, and Ido Dagan. 2015. “Improving Distributional Similarity with Lessons Learned from Word Embeddings.” *Transactions of the Association for Computational Linguistics* 3:211–25.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. “Efficient Estimation of Word Representations in Vector Space.” Retrieved 1 Jun 2018 (arXiv:1301.3781).