Language use shapes cultural stereotypes: Large scale evidence from gender

Supplementary Materials

Molly Lewis and Gary Lupyan

2019-03-15

Contents

This document was created from an R markdown file. The manuscript itself was also produced from an R markdown file, and all analyses presented in the paper can be reproduced from that document (https://github.com/mllewis/IATLANG/blob/master/writeup/journal/iat_lang.Rmd). The respository for the project can be found here: https://github.com/mllewis/IATLANG/.

NOTE: The SM is intended to be viewed interactively online at https://mlewis.shinyapps.io/iatlang_SI/.

Description of IAT data

As described in the Main Text, the IAT data come from Project Implicit (https://implicit.harvard.edu/implicit/; Nosek, Banaji, & Greenwald, 2002), for a sample collected 2005 - 2016.

Demographics

N by country

Number of participants by country after exclusions. Our final sample 657,335 participants from 39 countries. Participants were exclude who:

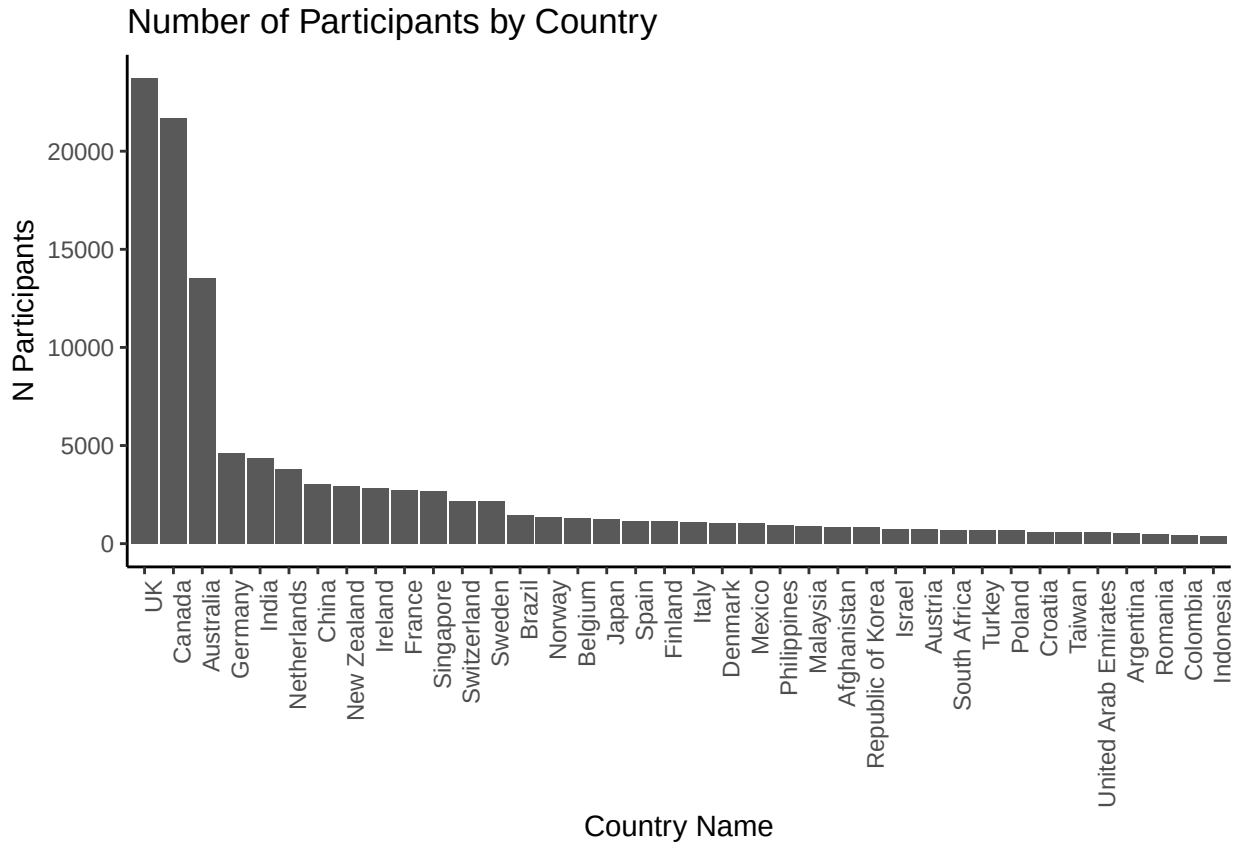
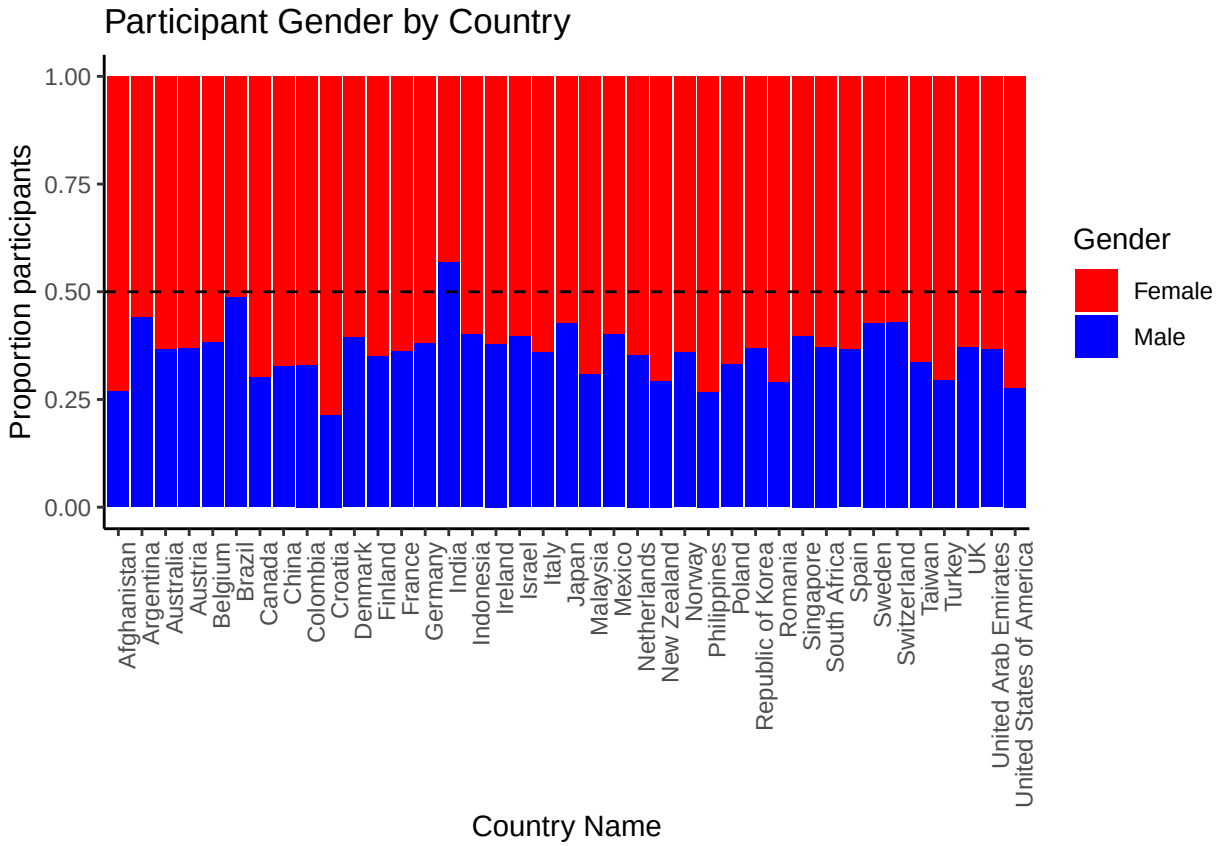


Figure 1: Note: Data from the US are excluded from this plot because of the large number of participants (N = 638,082).

- did not have complete gender, country, age, and implicit IAT measures (53%; the majority of these exclusions (69%) are due to missing IAT data - likely cases where the participant started but did not complete the IAT task).
- had average latencies for either critical block were over 1,800 ms or whose average overall latency was above 1,500 ms (as in Nosek, Banaji, & Greenwald, 2002; 5% of participants with complete data).
- made excess of 25% errors in any single critical block (as in Nosek, Banaji, & Greenwald, 2002; 14% of participants with complete data).
- were from countries with less than 400 participants (1% of remaining participants; in the "Correlations by language exclusion threshold" section below we show analyses with a range of threshold values)

Gender by country

Across countries, there tended to be more female participants, compared to male participants ($M = 0.36$ proportion males; $SD = 0.06$)



Age by country

Language by country

For each country, we identified the language with the most speakers using Ethnologue (Simons & Charles, 2018). Note that Ethnologue reports "Bavarian" as the primary language of Germany, and "Daric" as the primary language of Afghanistan. In order to map between the other data sources in our study, we used the more general language variant for these countries, German and Persian, respectively.

NOTE: See online version for this content (https://mlewis.shinyapps.io/iatlang_SI/)

Asterisks correspond to languages that were excluded from our analysis because word embedding models were unavailable.

Dependent Measures

Below are histograms for the implicit and explicit measures in the IAT Project Implicit data presented for each country separately. The implicit raw scores are the D-score values (estimate of gender bias, with postive values indicating strong bias to associate men with career), and the residualized values are the D-scores with participant age, participant sex and trial order residualized out. For the explicit measure, the raw score is the difference between participants answer to the question, "How strongly do you associate the following with males and females?" for the words "career" and for the word "family". Participants indicated their response on a Likert scale ranging from female (1) to male (7). For each participant, a single explicit score was calculated as the Career response minus the Family response, such that greater values indicate a greater bias to associate males with family. The residualized explicit value is the difference score with participant age,

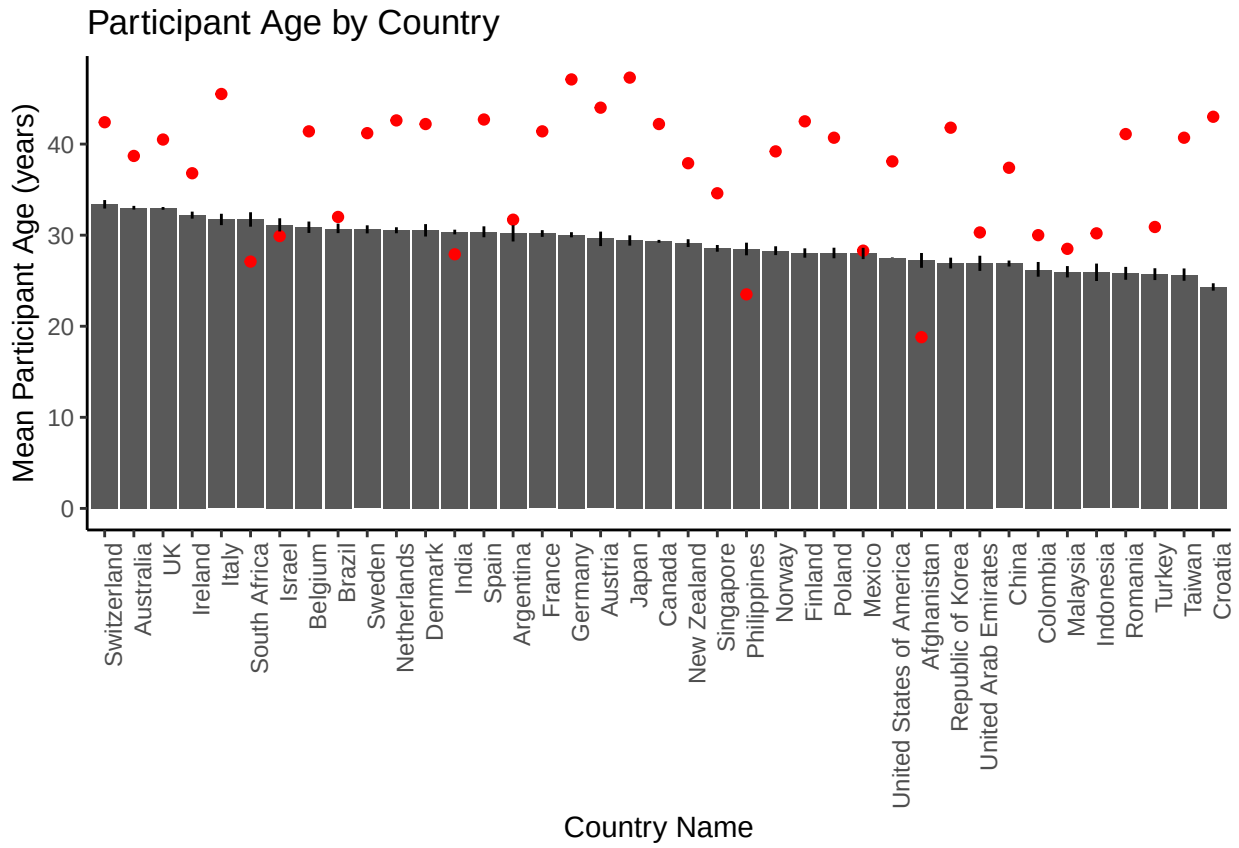
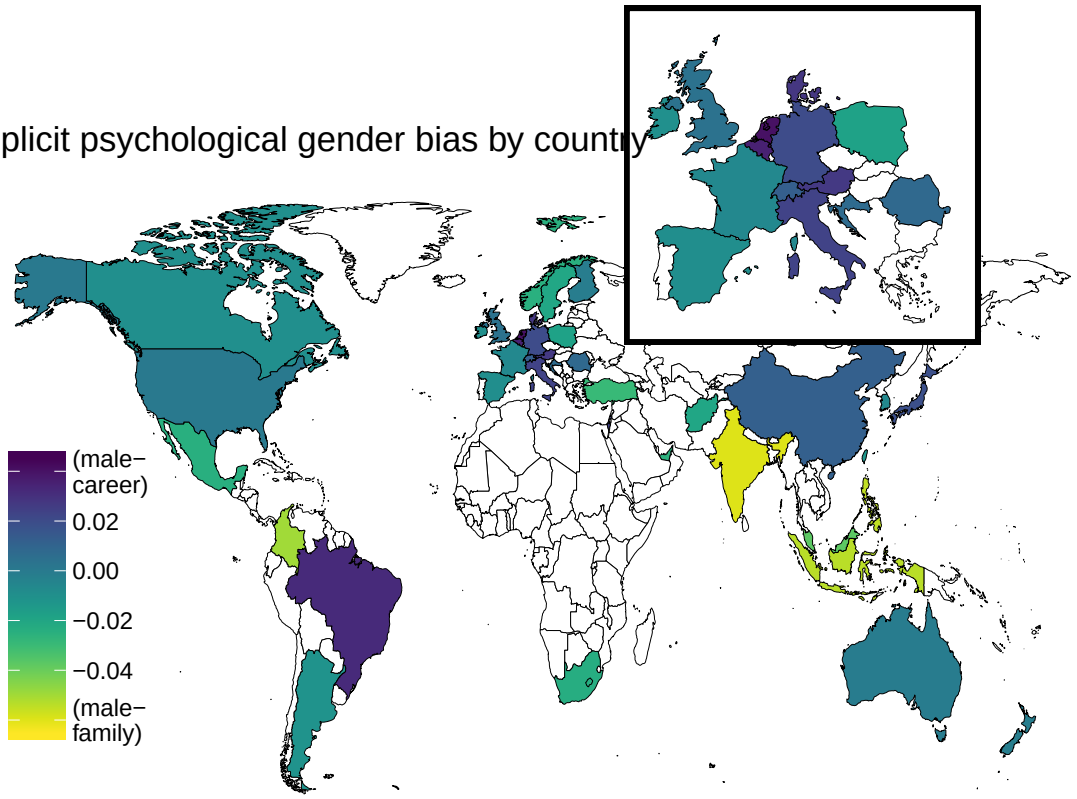


Figure 2: Bars show mean participant age by country; ranges correspond to 95% CIs. Red points show median age by country from CIA factbook data.

participant sex and trial order residualized out (we had no a priori reason for residualizing out trial order for explicit responses but did so to remain consistent with the residualized implicit measure). [see online version for this content; https://mlewis.shinyapps.io/iatlang_SI/]

Geographic distribution of IAT scores

Residualized implicit gender bias (IAT score) shown by country. Larger values (blue) indicate a larger bias to associate men with the concept of career and women with the concept of family. Countries in grey correspond to countries for which there was insufficient data to estimate the country-level gender bias. Inset shows IAT scores for European countries only.



Note that while Hindi is identified as the most frequently spoken language in India, India is highly multilingual and so Hindi embeddings may be a poor representation of the linguistic statistics for speakers inIndia as a group.

Replication of Caliskan et al. (2017)

Here we replicate the original set of Caliskan, Bryson, and Narayanan (2017; henceforth *CBN*) findings using the models trained on the corpora used in our paper, English Wikipedia (Bojanowski, Grave, Joulin, & Mikolov, 2016) and Subtitles (Lison & Tiedemann, 2016; Van Paridon & Thompson, in prep.).

For both the Wikipedia and Subtitle trained models, we calculate an effect size for each of the 10 biases reported in CBN which correspond to behavioral IAT results existing in the literature: flowers/insects–pleasant/unpleasant, instruments/weapons–pleasant/unpleasant, European-American/Afro-American–pleasant/unpleasant, males/females–career/family, math/arts–male/female, science/arts–male/female, mental-disease/physical-disease–permanent/temporary, and young/old–pleasant/unpleasant (labeled as Word-Embedding Association Test (WEAT) 1-10 in CBN). Note that CBN test three versions of race bias.

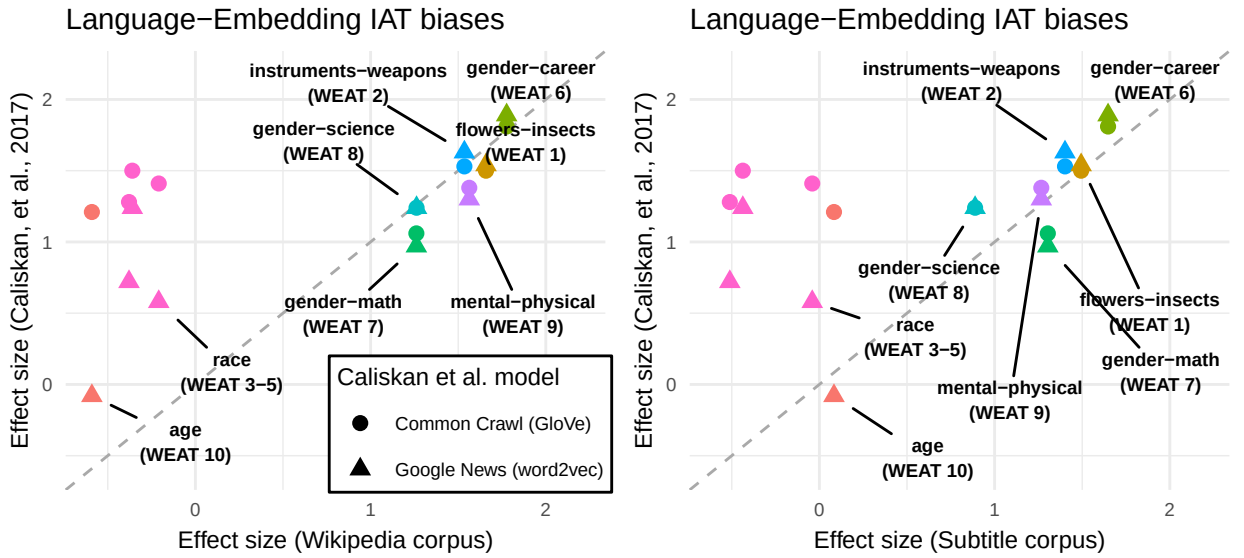


Figure 3: Effect sizes for the 10 IAT biases types (WEAT 1-10) reported in Caliskan et al. (2017; CBN). CBN effect sizes are plotted against effect sizes derived from the Wikipedia (left) and Subtitle (right) corpora. Point color corresponds to bias type, and point shape corresponds to the two CBN models trained on different corpora and with different algorithms.

We calculate the bias using the same effect size metric described in CBN, a standardized difference score of the relative similarity of the target words to the target attributes (i.e. relative similarity of male to career vs. relative similarity of female to career). This measure is analogous to the behavioral effect size measure where larger values indicate larger gender bias.

The figure below shows the effect size measures derived from the English Wikipedia corpus and the English Subtitle corpus plotted against effect size estimates reported by CBN from two different models (trained on the Common Crawl and Google News corpora). With the exception of biases related to race and age, effect sizes from our corpora are comparable to those reported by CBN. In particular, for the gender-career IAT—the bias relevant to our current purposes—we estimate the effect size to be 1.78, while CBN estimates it as approximately 1.85.

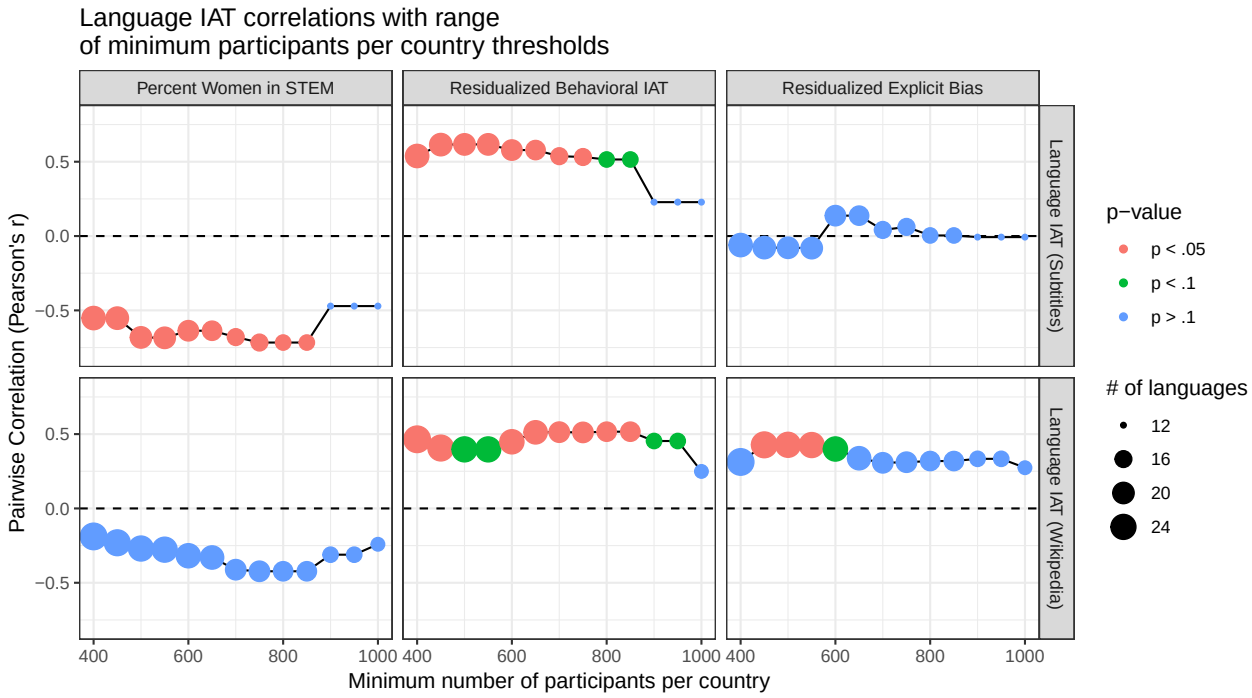
Descriptive statistics for all language-level measures

Below are the mean and standard deviations estimates for all measures presented in Table 1 of the Main Text.

Measure	Mean	SD
Residualized Explicit Bias	-0.009	0.177
Residualized Implicit Bias (IAT)	-0.008	0.027
Percent Women in STEM	14.876	4.462
Language IAT (Subtitle)	0.417	0.379
Language IAT (Wikipedia)	0.329	0.382
Prop. Gender-Distinct Labels	0.342	0.358
Occupation Bias (Subtitle)	0.017	0.024
Occupation Bias (Wikipedia)	0.030	0.029
Median Country Age	36.201	7.539

Correlations by language exclusion threshold

In the Main Text, we report psychological IAT data for participants who came from countries with at least 400 participants. This cutoff was largely arbitrary, but was selected to allow for a relatively large number of languages to be included in our analysis while also excluding languages with small sample sizes (and therefore less reliable estimates). Nevertheless, the pattern of findings we report in the Main Text remains



Study 2 language data

NOTE: See online version for this content (https://mlewis.shinyapps.io/iatlang_SI/).

Presented below are grammatical gender coding for each language and occupation items with translations in each language.

Grammatical gender coding

Below is the binary coding (gender vs. no gender) of each language for a sex-based grammatical gender system, based on WALS (Dryer & Haspelmath, 2013) and other sources when information was not available from WALS.

Occupation Items

In Study 2, we selected 20 occupation labels from the set of items used in Misersky, et al. (2014). Listed below are the 20 items along with their perceived gender bias as reported in Misersky, et al. (2014; larger values indicate occupation is perceived to be more closely associated with women).

Occupation Translations

Below are the translations of the 20 occupation words for each of the 25 target langauges. “Translation ID” identifies the translation when multiple translations were provided for a given occupation/language.

Predicting bias with both types of language predictors (Study 2)

In this analysis, we predict the magnitude of implicit bias by language with an additive linear model. As predictors, we include proportion gender distinct labels, linguistic bias (as measured by word embeddings of the IAT words), and median country age. Model coefficients are shown below for models based on the Subtitle (top) and Wikipedia (bottom) corpora.

Subtitle Corpus:

term	estimate	std.error	statistic	p.value
(Intercept)	0.072	0.150	0.484	0.634
Prop. Gender-Distinct Labels	0.328	0.159	2.063	0.054
Language IAT (Subtitle)	0.269	0.160	1.681	0.110
Median Country Age	0.298	0.168	1.776	0.093

Wikipedia Corpus:

term	estimate	std.error	statistic	p.value
(Intercept)	0.000	0.136	0.000	1.000
Prop. Gender-Distinct Labels	0.379	0.148	2.552	0.019
Language IAT (Wikipedia)	0.300	0.142	2.109	0.047
Median Country Age	0.414	0.150	2.759	0.012

References

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2016). Enriching word vectors with subword information.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186.

Dryer, M. S., & Haspelmath, M. (Eds.). (2013). *WALS online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. Retrieved from http://wals.info/

Lison, P., & Tiedemann, J. (2016). OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*.

Misersky, J., Gygax, P. M., Canal, P., Gabriel, U., Garnham, A., Braun, F., . . . others. (2014). Norms on the gender perception of role nouns in Czech, English, French, German, Italian, Norwegian, and Slovak. *Behavior Research Methods*, *46*(3), 841–871.

Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dynamics: Theory, Research, and Practice*, *6*(1), 101.

Simons, G. F., & Charles, D. F. (Eds.). (2018). Ethnologue: Languages of the world. Dallas, Texas: Online version: http://www.ethnologue.com. SIL International.

Van Paridon, J., & Thompson, B. (in prep.). Sub2Vec: Word embeddings from OpenSubtitles in 62 languages.