Language use shapes cultural stereotypes: Large scale evidence from gender

Molly Lewis[1,2] & Gary Lupyan[1]

[1] University of Wisconsin-Madison

[2] University of Chicago

Author Note

Portions of this manuscript appeared in Lewis & Lupyan, 2018, Cog. Sci Proceedings.

Correspondence concerning this article should be addressed to Molly Lewis, . E-mail: mollyllewis@gmail.com

Abstract

Cultural stereotypes such as the idea that men are more suited for paid work while women for taking care of the home and family may contribute to gender imbalances in STEM fields (e.g., Leslie, Cimpian, Meyer, & Freeland, 2015) and other undesirable gender disparities. Here, we test the hypothesis that word co-occurrence statistics (e.g., the co-occurence of "nurse" with "she") play a causal role in the formation of the men-career/women-family stereotype. We use word embedding models to measure bias in the distributional statistics of 25 languages and find that languages with larger biases tend to have speakers with larger implicit biases (N = 657,335). These biases are further related to the extent that languages mark gender in their lexical forms (e.g., "waiter"/"waitress") hinting that linguistic biases may be causally related to biases shown in people's implicit judgments.

Keywords: cultural stereotypes, implicit association task (IAT), gender

Word count: X

Language use shapes cultural stereotypes: Large scale evidence from gender

Introduction

By the time they are two years old, children have begun to acquire the gender stereotypes in their culture (Gelman, Taylor, Nguyen, Leaper, & Bigler, 2004). These stereotypes can have undesirable real-world consequences. For example, in one study, girls, compared to boys, were less likely to think that girls are "brilliant" and also less likely to choose activities that were described as for children "who are very, very smart" (Bian, Leslie, & Cimpian, 2017). In the aggregate, these behavioral choices could lead to the observed lower rates of female participation in science, technology, engineering, and mathematics (STEM) fields (Ceci & Williams, 2011; Leslie, Cimpian, Meyer, & Freeland, 2015; Miller, Eagly, & Linn, 2015; Stoet & Geary, 2018). Given the potential downstream consequences of cultural stereotypes, it is important to understand how the stereotypes are formed.

We can distinguish between two major sources of information on which stereotypes may be based is linguistic information. One is direct experience. For example, one may observe that most nurses one encounters are women and most philosophers are men and conclude that women better suited for nursing and men for philosophy. Another, source of information is language. Even in the absence of any direct experience with nurses or philosophers, a learner may observe that language about nurses and philosophers is more associated with female and male contexts, respectively. These contexts include proper names, pronouns, grammatical markers (particularly for languages with grammatical gender), and higher-order associations (e.g., seemingly non-gendered words can become gendered by being associated with explicitly gendered contexts).

Past research shows that even young children are sensitive to gender information delivered through language. For example, young children perform worse in a game if they are told that an anonymous member of the opposite gender performed better than they did on a

previous round (Rhodes & Brickman, 2008), or merely told that the game is associated with a particular gender (Cimpian, Mu, & Erickson, 2012). Further, there is evidence that descriptions as minimal as a single word can influence children's stereotypes: In one study, children were more likely to infer that a novel skill is stereotypical of a gender if the skill was introduced to children with a generic as opposed a non-generic subject ("[Girls are/There is a girl who is] really good at a game called 'gorp"'; Cimpian & Markman, 2011). To the extent that language is a source of information for forming cultural stereotypes, two people with similar direct experiences, but different linguistic experiences may come to have different stereotypes.

A common way of quantifying cultural stereotypes in individuals is by using the *Implicit Association Test* (IAT; Greenwald, McGhee, & Schwartz, 1998) which uses differences in reaction time in associations between different concepts (e.g. male-related words and career-related words vs.family-related words). As it turns out, various biases studied using IATs can be predicted from the distributional statistics of language. Caliskan, Bryson, and Narayanan (2017; henceforth *CBN*) measured the distance in semantic space between the words presented to participants in the IAT task. CBN found that these distances were highly correlated with the biases computed by a variety of IATs (e.g., valence and Caucasian vs. African-American names; gender and math-arts; mental vs. physical diseases and permanence). CBN only measured semantic biases in English. Here, we extend CBN's method to 25 languages examining whether languages with a stronger gender bias as expressed in distributional semantics predict stronger implicit and explicit gender biases on a large dataset of previously administered gender-career IAT (N = 657,335; Nosek, et al., 2002).

Discovering that stronger biases in language correlate with stronger biases in people's behavior can be interpreted in two ways. The first is that language merely *reflects* people's biases which are learned chiefly through nonlinguistic experiences. We refer to this as the

language as reflection hypothesis. The second possibility is that language exerts a causal influence on people's biases. We refer to this as the *language as causal agent* hypothesis. Work showing that linguistic descriptions can impact stereotypes in-the-moment (Cimpian & Markman, 2011; Cimpian et al., 2012; Rhodes & Brickman, 2008) already suggests that language *can* have some causal impact in an experimental context. We were interested in whether it actually does, and by what means.

Languages convey gender information in multiple ways including gender-specific proper names, pronouns, and titles (e.g., waiter vs. waitress). In addition, approximately 32% of languages have some type of grammatical gender (Dryer & Haspelmath, 2013) in which morphological markers are used to signal gender. Some previous work has shown that gender information conveyed by such morphological markers—which are also extended to inanimate objects—can infuse the inanimate objects with "natural" gender (Phillips & Boroditsky, 2003; Sera, Berge, & Castillo Pintado, 1994). More central to the goals of this project, because grammatical gender markers enter into agreement patterns, their use can exaggerate the extent to which gender is being communicated. For example, in Spanish the gender of a nurse has to be signaled grammatically: "enfermer*a*" vs. "enfermer*o*". In Russian, verbs (in some tenses) are inflected based on the gender of the speaker: "Ya ustal" (I-MASC am tired), but "Ya ustala" (I-FEM am tired). In both cases, it is not possible to leave the gender unspecified.

In Study 1a, we examine whether different languages convey gender roles differently in their distributional structure. That is, our analysis is not concerned with *what* is being said about typical gender roles, but rather the ways that *potentially* gendered words may be gendered by the morphological rules and statistical structure of different languages. In Study 1b we use this information to predict responses on a gender-career IAT. In Study 2 we examine whether the extent to which different languages use different forms for occupation terms (e.g., "waiter"/"waitress" but "teacher"/"teacher") correlates with greater implicit

gender bias thereby helping to narrow down the source of linguistic knowledge that may be playing a role in shaping gender stereotypes. Together, our data suggest that language not only reflects existing biases, but likely plays a causal role in shaping culturally-specific notions of gender.

Description of Cross-Cultural Dataset of Psychological Gender Bias

Materials and Methods

To quantify cross-cultural gender bias, we used data from a large-scale administration of an Implicit Association Task (IAT; Greenwald et al., 1998) by Project Implicit (https://implicit.harvard.edu/implicit/; Nosek, Banaji, & Greenwald, 2002). The IAT measures the strength of respondents' implicit associations between two pairs of concepts (e.g., male-career/female-family vs. male-family/female-career) accessed via words (e.g., "man," "business"). The underlying assumption of the IAT is that words denoting more similar meanings should be easier to pair together compared to more dissimilar pairs.

Meanings are paired in the task by assigning them to the same response keys in a two-alternative forced-choice categorization task. In the critical blocks of the task, meanings are assigned to keys in a way that is either bias-congruent (i.e. Key A = male/career; Key B = female/family) or bias-incongruent (i.e. Key A = male/family; Key B = female/career). Participants are then presented with a word related to one of the four concepts and asked to classify it as quickly as possible (see Study 1b Methods for list of target words). Slower reaction times in the bias-incongruent blocks relative to the bias-congruent blocks are interpreted as indicating an implicit association between the corresponding concepts (i.e. a bias to associate male with career and female with family).

We analyzed gender-career IAT scores collected by Project Implicit between 2005 and 2016, restricting our sample based on participants' reaction times and error rates using the

same criteria described in Nosek, Banjai, and Greenwald (2002, pg. 104). We only analyzed data for countries that had complete demographic information and complete data from the IAT for least 400 participants (2% of these respondents did not give responses to the explicit bias question). This cutoff was arbitrary, but the pattern of findings reported here holds for a range of minimum participant values (see SM[1]). Our final sample included 657,335 participants from 39 countries, with a median of 1,145 participants per country. Importantly, although the respondents were from largely non-English speaking countries, the IAT was conducted in English. We do not have language background data from the participants, but we assume that a large fraction of the respondents from non-English speaking countries were native speakers of the dominant language of the country and second language speakers of English. The fact that the test was administered in English lowers the prior likelihood of finding language-specific predictors of the kind we report here.

To quantify participants performance on the IAT we adopt the widely used *D-score*, which measures the difference between critical blocks for each participant while controlling for individual differences in response time (Greenwald, Nosek, & Banaji, 2003). After completing the IAT, participants were asked "How strongly do you associate the following with males and females?" for both the words "career" and "family." Participants indicated their response on a Likert scale ranging from *female* (1) to *male* (7). We calculated an explicit gender/career bias score for each participant as the Career response minus the Family response, such that greater values indicate a greater bias to associate males with career.

Results

At the participant level, implicit bias scores were positively correlated with participant age ($r = 0.06$, $p < .0001$). Male participants ($M = 0.32$, $SD = 0.37$) had a significantly smaller implicit gender bias than female participants ($M = 0.41$, $SD = 0.35$; $t = 96.82$, $p <$

[1]Available here: https://mollylewis.shinyapps.io/iatlang_SI/

.0001), a pattern consistent with previous findings (Nosek et al., 2002). Finally, implicit bias scores were larger for participants that received the block of trials with bias-incongruent mappings first relative to the opposite order (t = -104.03, $p < .0001$).

Because we did not have language information at the participant level, in the remaining analyses we examine gender bias and its predictors at the country level. To account for the above-mentioned influences on implicit bias, we calculated a residual implicit bias score for each participant, controlling for participant age, participant sex, and block order. We also calculated a residual explicit bias score controlling for the same set of variables. We then averaged across participants to estimate the country-level gender bias (implicit: M = -0.01; SD = 0.03; explicit: M = 0.00; SD = 0.18). Implicit gender biases were moderately correlated with explicit gender biases at the level of participants (r = 0.16, $p < .0001$) but not countries (r = 0.26, p = 0.11).

Do the implicit and explicit biases measured by the Project Implicit dataset predict any real world outcomes? We compared our residual country-level implicit and explicit gender biases to a gender equality metric reported by the United Nations Educational, Scientific and Cultural Organization (UNESCO) for each country: the percentage of women among STEM graduates in tertiary education from 2012 to 2017 (Miller et al., 2015; Stoet & Geary, 2018; available here: http://data.uis.unesco.org/). These data were available for 33 out of 39 of the countries in our sample. Consistent with previous research (Miller et al., 2015), we found that implicit gender bias was negatively correlated with percentage of women in STEM fields: Countries with a smaller gender bias tended to have more women in STEM fields (r = -0.54, $p < .001$). In contrast, there was no relationship between the percentage of women in STEM fields and the explicit gender-bias measure used by Project Implicit (r = 0.14, p = 0.43). In addition, we found a strong correlation between the median age of each country's population (as reported by the CIA factbook, 2017) and the residual implicit bias (in which participant age was held constant): Countries with older populations

tended to have larger gender biases ($r = 0.63$, $p < .0001$).

In sum, we replicate previously-reported patterns of gender bias in the gender-career IAT literature, with roughly comparable effect sizes (c.f. Nosek, et al., 2002). The weak correlation between explicit and implicit measures is consistent with claims that these two measures tap into different cognitive constructs (Forscher et al., 2016). In addition, we find that an objective measure of gender equality—female enrollment in STEM fields—is associated with implicit gender bias. The finding that older participants show stronger biases may stem from a cohort effect, but it is not obvious why there is a strong positive association between the median age of a country's population and a larger implicit bias when adjusting for the age of individual participants.

Study 1: Relating gender biases in distributional semantics and human behavior

Are participants' gender biases predictable from the language to which they are exposed? For example, are the semantics of the words "woman" and "family" more similar in Spanish than in English? Both the language-as-reflection and language-as-causal-factor hypotheses predict a positive correlation between the measured biases and biases present in language.

General approach

To model word meanings, we use semantic embeddings derived from model that learns meanings by trying to predict a word from surrounding words, given a large corpus. The core assumption of these models is that the meaning of a word can be described by the words it tends to co-occur with—words occurring in similar contexts, tend to have similar meanings (Firth, 1957). A word like "dog", for example is represented as more similar to

"cat" and "hound" than to "banana" because "dog" co-occurs with words more in common with "cat" and "hound" than with "banana" (Landauer & Dumais, 1997; Lund & Burgess, 1996). Recent developments in machine learning allow the idea of distributional semantics to be implemented in a way that takes into account many features of language structure while remaining computationally tractable. The best known of these word embedding models is *word2vec* (Mikolov, Chen, Corrado, & Dean, 2013). By attempting to predict the words that surround another word, the model is able to learn a vector-based representation for each word that represents its similarity to other words, i.e., a semantic embedding. We can then compute the similarity between two words by taking the distance between their vectors (e.g., cosine of angle).

We begin by validating word embedding measures of gender bias by comparing them to explicit human judgements of word genderness (Study 1a). We then apply this method to models trained on text in other languages (Study 1b). To foreshadow the results, we find that the implicit gender biases reported in Study 1 for individual countries are correlated with the biases found in the distributional semantics of the language spoken in the countries of the participants.

Study 1a: Word embeddings as a measure of psychological gender bias

Methods. To validate word embeddings as a measure of psychological gender bias, we asked whether words that were closely associated with males in the word embedding models tended to be rated by human participants as being more male biased. We found human and word-embedding estimates of gender bias to be highly correlated.

We used an existing set of word norms in which participants were asked to rate "the gender associated with each word" on a Likert scale ranging from *very feminine* (1) to *very masculine* (7; Scott, Keitel, Becirspahic, Yao, & Sereno, 2018). We compared these norms to

estimates of gender bias obtained from embedding models pre-trained on two different corpora of English text: Wikipedia (Bojanowski, Grave, Joulin, & Mikolov, 2016) and subtitles from movies and TV shows (Lison & Tiedemann, 2016; Van Paridon & Thompson, in prep.). The Wikipedia corpus is a large, naturalistic corpus of written language; the subtitle corpus is a smaller corpus of spoken language. Both models were trained using the fastText algorithm (a variant of word2vec; Joulin, Grave, Bojanowski, & Mikolov, 2016).

To calculate a gender score from the word embeddings, for each word we calculated the average cosine distance to a standard set of male "anchor" words: ("male," "man," "he," "boy," "his," "him," "son," and "brother") and the average cosine similarity to a set of female words ("female," "woman," "she," "girl," "hers," "her," "daughter," and "sister"). A gender score for each word was then obtained by taking the difference of the similarity estimates (mean male similarity - mean female similarity), such that larger values indicated a stronger association with males. There were 4,671 words in total that overlapped between the word-embedding models and human ratings.

Results. Estimates of gender bias from the Subtitle corpus ($M = 0.01$; $SD = 0.03$) and the Wikipedia corpus ($M = 0$; $SD = 0.03$) were highly correlated with each other ($r = 0.71$; $p < .0001$). Critically, bias estimates from both word embedding models were also highly correlated with human judgements ($M = 4.10$; $SD = 0.92$; $r_{\text{Subtitle}} = 0.63$; $p < .0001$; $r_{\text{Wikipedia}} = 0.59$; $p < .0001$; Fig. 1). This suggests that the psychological gender bias of a word can be reasonably estimated from word embeddings.

Study 1b: Gender bias across languages

Having validated our method, we next turn toward examining the relationship between psychological and linguistic gender biases. In Study 1b, we estimate the magnitude of the gender-career bias in the dominant language spoken in the countries of the Project Implicit

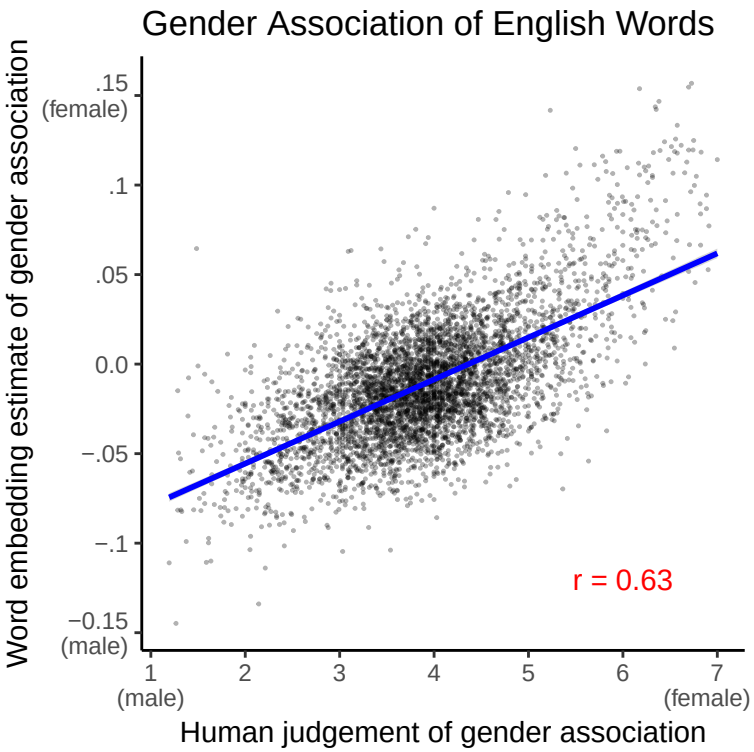

Figure 1. Word estimates of gender bias from the Subtitle-trained embedding model as a function of human judgments of gender bias (Study 1a). Each point corresponds to a word. Larger numbers indicate stronger association with females (note that this differs from the design of the rating task, but is changed here for consistency with other plots). Blue line shows linear fit and the error band band indicates standard error (too small to be visible).

participants and compare it with estimates of psychological gender bias from the Project Implicit data set.

Methods. We identified the most frequently spoken language in each country in our analysis using the Ethnologue (Simons & Charles, 2018). This included a total of 26 unique languages.[2] For each language, we obtained translations from native speakers for the stimuli

[2]Our final sample of languages (N = 25) excluded Zulu because word embedding measures were not available for this language (see below). Note also that while Hindi is identified as the most frequently spoken language in India, India is highly multi-lingual and so Hindi embeddings may be a poor representation of the linguistic statistics for speakers in India as a group

in the Project Implicit gender-career IAT behavioral task (Nosek et al., 2002) with one slight modification. In the behavioral task, proper names were used to cue the male and female categories (e.g. "John," "Amy"), but because there are not direct translation equivalents of proper names, we instead used a set of generic gendered words which had been previously used for a different version of the gender IAT (e.g., "man," "woman;" Nosek et al., 2002). Our linguistic stimuli were therefore a set of 8 female and 8 male Target Words (identical to Study 1a), and the set of 8 Attribute Words words used in the Project Implicit IATs: 8 related to careers ("career," "executive," "management," "professional," "corporation," "salary," "office," "business") and 8 related to families ("family," "home," "parents," "children," "cousins," "marriage," "wedding," "relatives"). For one language, Tagalog, we were unable to obtain translations from a native speaker, and so Tagalog translations were compiled from dictionaries.

We used these translations to calculate a gender bias effect size from word embedding models trained on text in each language. Our effect size measure is 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). Our effect size measure is identical to that used by CBN with an exception for grammatically gendered languages (see SM for replication of CBN on our corpora). Namely, for languages with grammatically gendered Attribute Words (e.g., niñas for female children in Spanish), we calculated the relationship between target words and attribute words of the same gender (i.e. "hombre" (man) to "niños" and "mujer" (woman) to "niñas"). In cases where there were multiple translations for a word, we averaged across words such that each of our target words was associated with a single vector in each language. In cases where the translation contained multiple words, we used the entry for the multiword phrase in the model when present, and averaged across words otherwise. Like the behavioral effect size from the Project Implicit data, larger values indicate larger gender bias.

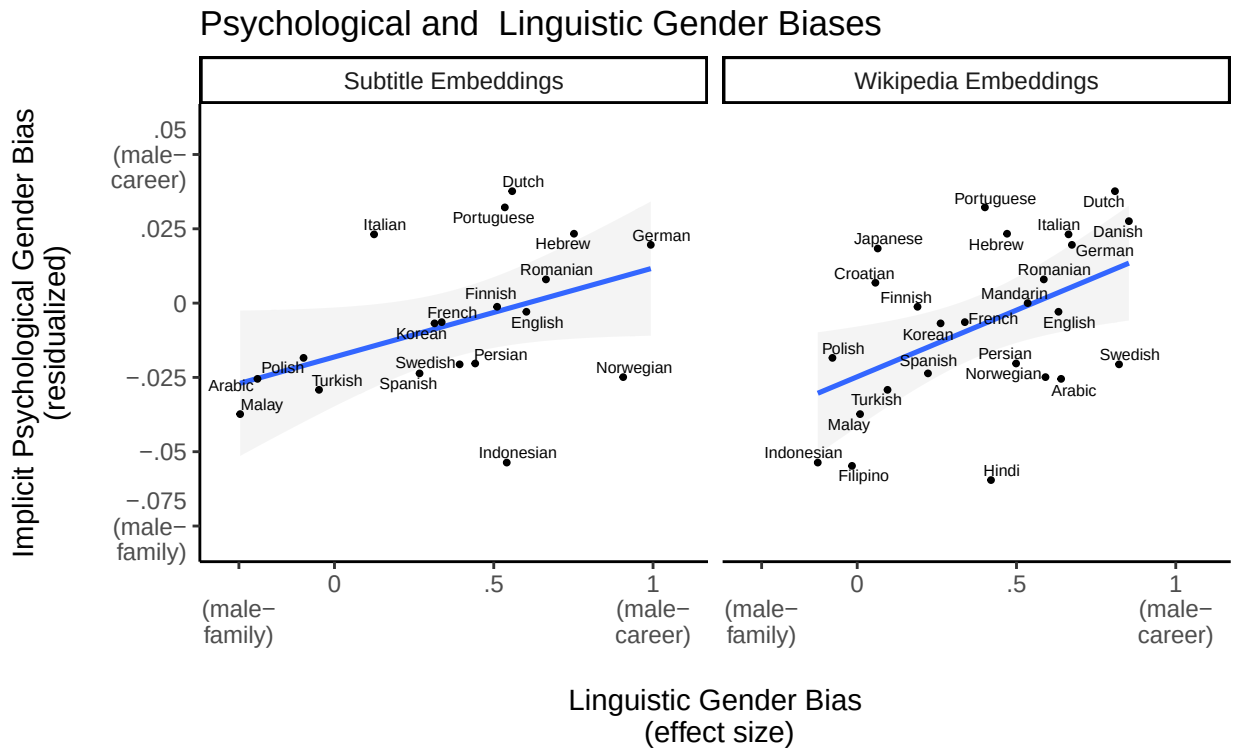

Figure 2. Implicit psychological gender bias (adjusted for age, sex, and block order) as a function of the linguistic gender bias. Each point corresponds to a language (Study 1b). Linguistic biases are estimated from models trained on text in each language from a Subtitle corpus (left) and a Wikipedia corpus (right). Larger values indicate a larger bias to associate men with the concept of career and women with the concept of family. Error band indicates standard error of the linear model estimate

We calculated gender bias estimates using the same word-embedding models as in Study 1a (Subtitle and Wikipedia corpora). We excluded languages from the analysis for which 20% or more of the target words were missing from the model or the model did not exist (see SM for more details). This led us to exclude one language (Zulu) from the analysis of the Wikipedia corpus and six languages from the analysis of the Subtitle corpus (Chinese, Croatian, Hindi, Japanese, Tagalog, and Zulu). Our final sample included 25 languages in total ($N_{\text{Wikipedia}} = 25$; $N_{\text{Subtitle}} = 20$), representing 8 language families.

Results. Despite the differences in the specific content conveyed by the Subtitles and Wikipedia corpus, the estimated gender bias for each language was similiar across the two corpora ($t(19)$ = -0.06, p = 0.95). We next examined the relationship between these estimates of linguistic gender bias and the bias scores on the IAT of participants from countries where that language was dominant (and, we assume, was the native language of most of these individuals). Implicit gender bias was positively correlated with estimates of language bias from both Subtitle () and Wikipedia (; Fig. 2; Table 1 shows the language-level correlations between all variables in Studies 1b and 2). The relationship between implicit gender bias and language bias remained reliable after partialling out the effect of median country age (Subtitle: ; Wikipedia:). Linguistic gender bias was not correlated with explicit gender bias (Subtitle: r = -0.08, p = 0.74; Wikipedia: r = 0.34, p = 0.09). Estimates of language bias from the Subtitle corpus were correlated with the objective measure of gender equality, percentage of women in STEM fields (r = -0.55, p = 0.02); this relationship was not reliable for the Wikipedia corpus (r = -0.19, p = 0.4).

Table 1

Correlation (Pearson's r) for all measures in Study 1 and 2 at the level of languages. Numbers in parentheses show partial correlations controlling for median country age. Single astericks indicate $p < .05$ and double astericks indicate $p < .01$. The + symbol indicates a marginally significant p-value, $p < .1$.

	Residualized Explicit Bias	Residualized Implicit Bias (IAT)	Percent Women in STEM	Language IAT (Subtitle)	Language IAT (Wikipedia)	Occupation Bias (Subtitle)	Occupation Bias (Wikipedia)	Prop. Gender-Distinct Labels	Median Country Age
Residualized Explicit Bias		.18 (.28)	.18 (.17)	-0.08 (-0.06)	.34+ (.38+)	.28 (.33)	.29 (.33)	.1 (.14)	-0.07
Residualized Implicit Bias (IAT)	.18 (.28)		-0.52* (-0.37+)	.5* (.42*)	.48* (.43*)	.64** (.57**)	.59** (.51*)	.56** (.47*)	.61**
Percent Women in STEM	.18 (.17)	-0.52* (-0.37+)		-0.55* (-0.49*)	-0.19 (-0.09)	-0.39 (-0.28)	-0.32 (-0.2)	-0.35 (-0.23)	-0.42+
Language IAT (Subtitle)	-0.08 (-0.06)	.5* (.42*)	-0.55* (-0.49*)		.51* (.47*)	.42+ (.35+)	.4+ (.33)	.28 (.2)	.31
Language IAT (Wikipedia)	.34+ (.38+)	.48* (.43*)	-0.19 (-0.09)	.51* (.47*)		.28 (.21)	.44* (.39+)	.18 (.11)	.25
Occupation Bias (Subtitle)	.28 (.33)	.64** (.57**)	-0.39 (-0.28)	.42+ (.35+)	.28 (.21)		.8** (.77**)	.75** (.71**)	.36
Occupation Bias (Wikipedia)	.29 (.33)	.59** (.51*)	-0.32 (-0.2)	.4+ (.33)	.44* (.39+)	.8** (.77**)		.7** (.66**)	.34+
Prop. Gender-Distinct Labels	.1 (.14)	.56** (.47*)	-0.35 (-0.23)	.28 (.2)	.18 (.11)	.75** (.71**)	.7** (.66**)		.35+
Median Country Age	-0.07	.61**	-0.42+	.31	.25	.36	.34+	.35+	

Discussion. In Study 1, we found that that a previously reported psychological gender bias – the bias to associate men with career and women with family – was correlated with the magnitude of that same bias as measured in language statistics of 25 languages. Participants completing the IAT in countries where the dominant language had stronger associations between men and career words, and women and family words, showed stronger biases on the gender-career IAT. This result is consistent with both the *language-as-reflection* and *language-as-causal-factor* hypotheses. In Study 2, we try to better distinguish between the two hypotheses by investigating whether the gender-career bias is associated with two structural features of language: grammatical gender, and the presence of gendered occupation terms (e.g., waiter/waitress).

Study 2: Gender bias and lexicalized gender

Languages can mark gender distinctions on words referring to people either as part of a grammatical gender system, as in Spanish ("enfermero" (nurse-MASC) versus "enfermera" (nurse-FEM)) and through idiosyncratic marking on particular nouns (e.g., "waiter" versus "waitress" in English). Languages that have grammatical gender systems make gender distinctions more frequently, since it is an obligatory part of the grammar. Further, languages that mark gender on the noun tend to mark gender on other arguments in the sentence as part of an agreement system ("el enfermo alto" (the tall nurse-MASC) versus "la enferma alta" (the tall nurse-FEM)), potentially making the reference to biological sex even more salient to speakers.

In Study 2, we examined whether grammatical gender and preponderance of gender-specific occupation terms are associated with a greater gender-career bias as measured by the IAT and whether this relationship is further mediated by the language statistics. Finding such associations would lend support to the language-as-causal-agent hypothesis because grammatical gender and (to a lesser degree) lexical gender encoding are

relatively stable features of language. Although both are subject to change over time, these changes are somewhat independent of the propositional content conveyed by language. For example, a Finnish document about nursing being unsuitable for men would still use a gender-neutral form of "nurse" while a Spanish document promoting nursing careers to men would be comitted to using a gender-specific form of "nurse" throughout.

Methods. We identified 20 occupation names that were likely to have corresponding terms in all 25 of our languages, and that were balanced in terms of their perceived gender bias in the workforce (Misersky et al., 2014). We then translated these words into each of the 25 languages in our sample, distinguishing between male and female variants (e.g., "waiter" vs. "waitress") where present. The words were translated by consulting native speakers and dictionaries, as necessary.

To estimate the extent to which a language lexically encoded gender, we calculated the proportion of occupations within each language for which the male and female forms differed. Larger values indicate a preponderance for more gender-specific forms in a language. We also estimated the extent to which each occupation term was gender biased in its language statistics using word embedding models trained in each language on the Subtitle and Wikipedia corpora. For each occupation form, we estimated its bias in language statistics using the same pairwise similarity metric as in Study 1a, and then averaged across occupations within a language to get a language-level estimate of gender bias in language statistics. Larger values indicate greater gender bias in language statistics. Finally, we coded each language for the presence or absense of a sex-based grammatical gender system using WALS (Dryer & Haspelmath, 2013) and consulting other sources, as necessary. We then compared each of the three language measures (proportion specific gender forms, bias in language statistics, and grammatical gender) to the psychological gender measures described in Study 1b (implicit association bias effect size and explicit bias, adjusted for age, sex and block order).

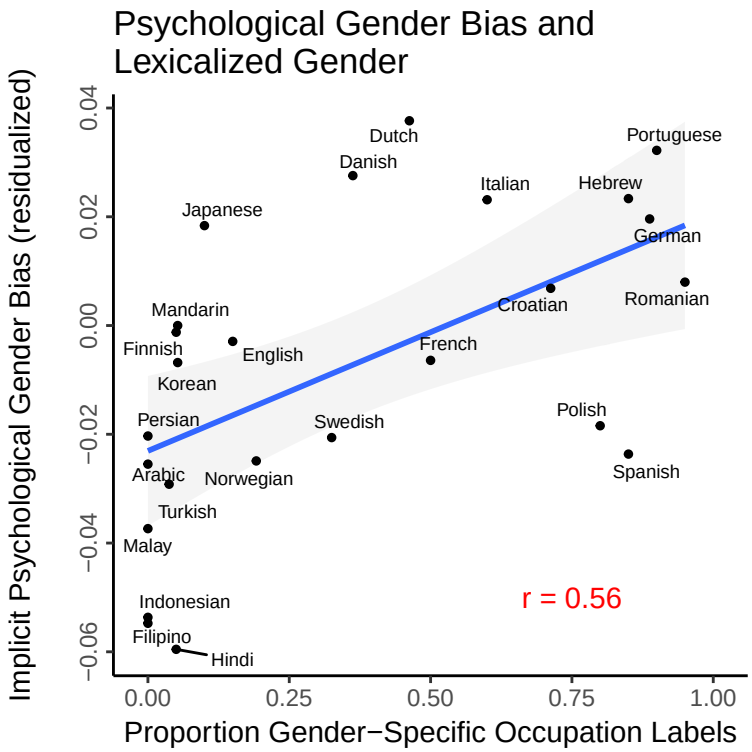

Figure 3. Residualized implicit psychological gender bias as a function of the proportion of gender-specific labels for set of words referring to occupations. Error band indicates standard error of the linear model estimate.

Results. In an additive linear model controling for median country age, there was no difference in implict or explicit psychological gender bias for speakers of languages with a grammatical gender system (N = 12), compared to those without (N = 13; Implicit: β = 0; SE = 0.01; t = -0.4; Explicit: β = -0.09; SE = 0.07; t = -1.21). However, degree of implicit psychological gender bias was systematically related to lexical marking on occupation words: Languages with more gender-specific forms tended to have speakers with higher psychological gender bias (M = 0.36, SD = 0.36; r = 0.56, $p < .01$), even after partialling out the effect of median country age (r = 0.47, p = 0.02; Table 1). There was no relationship between explicit psychological gender bias and lexical marking of occupation words after partialling out the effect of median country age (r = 0.14, p = 0.52).

We next examined whether having gender-specific forms for a particular occupation

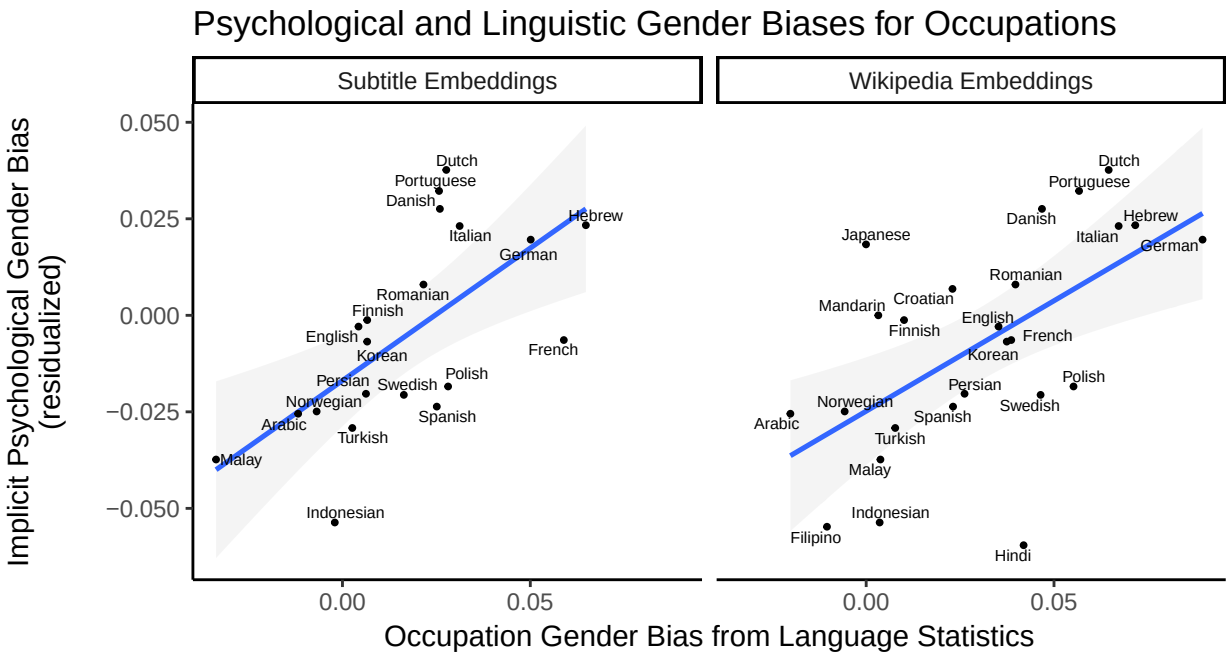

Figure 4. Residualized psychological IAT gender bias as a function of mean gender bias of words referring to occupations, with each point corresponding to a language (Study 2). Linguistic biases are estimated from models trained on text in each language from a Subtitle corpus (left) and a Wikipedia corpus (right). Error bands indicate standard error of the linear model estimate.

was associated with greater gender bias in the language statistics for that form. We fit a mixed effect model predicting degree of gender bias in language statistics (estimated from word embedding models) as a function of degree of distinctiveness between male and female forms for that word, with random intercepts and slopes by language. Degree of form distinctiveness was a strong predictor of language statistics for models trained on both the Subtitle corpus ($\beta = 0.59$; $SE = 0.07$; $t = 8.72$) and Wikipedia corpus ($\beta = 0.81$; $SE = 0.09$; $t = 9.48$), with words with shared male and female forms tending to have less gender bias. This relationship also held at the level of languages: Languages with more distinct forms had a greater gender-career bias in language statistics (Subtitle corpus: $r = 0.75$, $p < .01$; Wikipedia corpus: $r = 0.7$, $p < .01$).

Finally, we examined the relationship between gender bias in language statistics and psychological gender biases at the level of languages. Unlike in Study 1, all the target words in the present study referred to people (occupations) and thus potentially could be marked for the gender of the referenced person. Consequently, if explicit gender marking drives language statistics, we should expect to see a strong positive relationship at the level of languages between bias in language statistics *for occupation words* and psychological biases for speakers of that language. Consistent with this prediction, gender bias in language statistics for occupation words were positively correlated with implicit psychological gender bias for both language models (Subtitle: $r = 0.64$, $p < .01$; Wikipedia: $r = 0.59$, $p < .01$), and remained reliable after partialling out the effect of median country age (Subtitle: $r = 0.57$, $p < .01$; Wikipedia: $r = 0.51$, $p = 0.01$; Figure 4). There was no relationship between language statistics for occupation words and explicit psychological gender bias, even after partialling out the effect of median country age (Subtitle: $r = 0.33$, $p = 0.12$; Wikipedia: $r = 0.33$, $p = 0.11$).

To understand the relative predictive power of language statistics and distinct forms, we fit an additive linear model predicting implicit psychological bias from language statistics and proportion distinct forms, controlling for median country age. Neither bias in language statistics (Wikipedia: $\beta = 0.3$, $SE = 0.21$, $Z = 1.46$, $p = 0.16$; Subtitle: $\beta = 0.37$, $SE = 0.25$, $Z = 1.45$, $p = 0.17$) nor proportion distinct forms (Wikipedia: $\beta = 0.2$, $SE = 0.21$, $Z = 0.98$, $p = 0.34$; Subtitle: $\beta = 0.18$, $SE = 0.25$, $Z = 0.74$, $p = 0.47$) predicted implicit bias, likely due to collinearity between the two measures (Wikipedia: $r = 0.70$, $p < .001$; Subtitle: $r = 0.75$, $p < .001$; VIF >= 2 for language statistics and proportion distinct forms in both Subtitle and Wikipedia models). The high degree of collinearity between language statistics and proportion distinct forms is consistent with a causal model in which language statistics mediate the effect of distinct gender forms on implicit psychological bias: The presence of distinct forms referring to people of different genders *leads to* biased language statistics, which in turn leads to gender bias in behavior. Consistent with this model, a bootstrap test

of mediation revealed a marginal effect for the Subtitle model (path-ab = 0.28, $p = 0.10$; Alfons, 2018), and significant mediation effect for the Wikipedia model (path-ab = 0.34, $p = 0.04$).[3]

Discussion

General Discussion

Across two studies, we sought to understand the role that language plays in the formation of one particular gender stereotype: the bias to associate men with career and women with family. We estimated the magnitude of this bias among speakers of 25 different languages using a large scale administration of the IAT, and then examined how the strength of the stereotype related to gender biases found in language. In Study 1, we found that languages that have a higher degree of gender bias in their co-occurrence statistics tend to have speakers that have stronger gender stereotypes. In Study 2, we found that words that were explicitly marked for the gender of the referent ("waiter"/"waitress") tended to have more biased statistics, and that languages that more frequently explicitly marked the gender of referents tended to have speakers that were overall more biased. Together these studies provide evidence that the statistics of language use may have a causal influence on the formation of cultural stereotypes.

Our work is the first to characterize the relationship between broad structural patterns in language and cultural stereotypes. Broadly, the positive correlation that we find between these two variables – gender bias in language and gender bias in speakers – is consistent with both the possibility that language plays a causal role in the emergence of cultural stereotypes and the idea that language merely reflects existing stereotypes in its speakers.

[3]Though our power to detect this effect is relatively low (approximately, .4; Schoemann, Boulton, & Short, 2017).

Study 2 however provides suggestive evidence for the causal hypothesis: We find that X. Importantly these two causal forces are not mutually exclusive, and in fact may amplify the effects of each other. For example, the fact that nurses tend to be female in the workforce likely leads speakers to talk about women as nurses more often than men, which in turn produces biased language statistics that may ultimately serve to reinforce a pre-existing cultural stereotype. Future work could use experimental methods to manipulate language statistics in order to more directly examine these causal influences.

The implications of the language-as-causal-factor hypothesis are important: The mere process of listening to and producing language exposes one to statistics that may lead to the formation of cultural stereotypes. Given that the average adult produces ~16,000 words per day (Mehl, Vazire, Ramírez-Esparza, Slatcher, & Pennebaker, 2007), the aggregate effects of exposure to biased language statistics could be large, particularly for children who are just beginning to develop gender stereotypes (Gelman et al., 2004). Further, the fact that these language statistics are not explicitly accessible to speakers (versus, for example, the statement, "Most nurses are women.") may make it harder to combat these stereotypes and ultimately exaggerate their effects (Silverstein, 1981). Future work is needed to understand the relative influence of more explicit messages of gender bias compared to the implicit biases present in language statistics.

An important limitation of the current work is that our measure of the psychological reality of the gender stereotype relies on the IAT, a measure that has been criticized for both its low reliability (Lane, Banaji, Nosek, & Greenwald, 2007) and limited validity (e.g., Oswald, Mitchell, Blanton, Jaccard, & Tetlock, 2013). Issues of reliability are less relevant here because we use the IAT to measure group-level differences, rather than as an individual-difference measure. However, issues of validity are an important concern, particularly since we find that languages estimates of bias are uncorrelated with explicit measures, althose these measures were extrmeely coarse (others have pointed to IAT's lack of

predictive validity; Fazio & Olson, 2003). Understanding the full import of linguistic biases on cultural stereotypes would therefore require obtaining measures more closely related to real-world behavior.

In conclusion, information about cultural stereotypes is necessarily acquired through experience in that culture. In our work here, we have suggested that the statistics of language use are an important element of that experience. Many cultural associations and stereotypes are present in the statistics of language may be innocuous – indeed, these statistics may be an important mechanism through which cultural information is transmitted. But, in other cases, like the kind of gender stereotypes investigated here, language may be playing a powerful role in their formation, and ultimately contributing to undesirable real-world consequences like gender inequality in STEM. Understanding the causal role that language plays in the formation of these stereotypes is therefore an important first step to changing these consequences.

References

Alfons, A. (2018). *Robmed: (Robust) mediation analysis*. Retrieved from https://CRAN.R-project.org/package=robmed

Bian, L., Leslie, S.-J., & Cimpian, A. (2017). Gender stereotypes about intellectual ability emerge early and influence children's interests. *Science*, *355*(6323), 389–391.

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.

Ceci, S. J., & Williams, W. M. (2011). Understanding current causes of women's underrepresentation in science. *Proceedings of the National Academy of Sciences*, 201014871.

Central Intelligence Agency (CIA). (2017). The World Factbook. Retrieved from https://www.cia.gov/library/publications/the-world-factbook/index.html

Cimpian, A., & Markman, E. M. (2011). The generic/nongeneric distinction influences how children interpret new information about social others. *Child Development*, *82*(2), 471–492.

Cimpian, A., Mu, Y., & Erickson, L. C. (2012). Who is good at this game? Linking an activity to a social category undermines children's achievement. *Psychological Science*, *23*(5), 533–541.

Dryer, M. S., & Haspelmath, M. (Eds.). (2013). *WALS online*. Leipzig: Max Planck

Institute for Evolutionary Anthropology. Retrieved from http://wals.info/

Fazio, R. H., & Olson, M. A. (2003). Implicit measures in social cognition research: Their meaning and use. *Annual Review of Psychology*, *54*(1), 297–327.

Firth, J. (1957). A synopsis of linguistic theory 1930-1955 in studies in linguistic analysis, philological society. Oxford.

Forscher, P. S., Lai, C., Axt, J., Ebersole, C. R., Herman, M., Devine, P. G., & Nosek, B. A. (2016). A meta-analysis of change in implicit bias.

Gelman, S. A., Taylor, M. G., Nguyen, S. P., Leaper, C., & Bigler, R. S. (2004). Mother-child conversations about gender: Understanding the acquisition of essentialist beliefs. *Monographs of the Society for Research in Child Development*, i–142.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, *74*(6), 1464.

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the Implicit Association Test: An improved scoring algorithm. *Journal of Personality and Social Psychology*, *85*(2), 197.

Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2016). Bag of tricks for efficient text classification. *arXiv Preprint arXiv:1607.01759*.

Landauer, T. K., & Dumais, S. T. (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211.

Lane, K. A., Banaji, M. R., Nosek, B. A., & Greenwald, A. G. (2007). Understanding and

using the implicit association test: IV. *Implicit Measures of Attitudes*, 59–102.

Leslie, S.-J., Cimpian, A., Meyer, M., & Freeland, E. (2015). Expectations of brilliance underlie gender distributions across academic disciplines. *Science*, *347*(6219), 262–265.

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*.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, *28*(2), 203–208.

Mehl, M. R., Vazire, S., Ramírez-Esparza, N., Slatcher, R. B., & Pennebaker, J. W. (2007). Are women really more talkative than men? *Science*, *317*(5834), 82–82.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space.

Miller, D. I., Eagly, A. H., & Linn, M. C. (2015). Women's representation in science predicts national gender-science stereotypes: Evidence from 66 nations. *Journal of Educational Psychology*, *107*(3), 631.

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.

Oswald, F. L., Mitchell, G., Blanton, H., Jaccard, J., & Tetlock, P. E. (2013). Predicting ethnic and racial discrimination: A meta-analysis of iat criterion studies. *Journal of Personality and Social Psychology*, *105*(2), 171.

Phillips, W., & Boroditsky, L. (2003). Can quirks of grammar affect the way you think? Grammatical gender and object concepts. In *Proceedings of the 25th Annual Meeting of the Cognitive Science Society* (pp. 928–933).

Rhodes, M., & Brickman, D. (2008). Preschoolers' responses to social comparisons involving relative failure. *Psychological Science*, *19*(10), 968–972.

Schoemann, A. M., Boulton, A. J., & Short, S. D. (2017). Determining power and sample size for simple and complex mediation models. *Social Psychological and Personality Science*, *8*(4), 379–386.

Scott, G. G., Keitel, A., Becirspahic, M., Yao, B., & Sereno, S. C. (2018). The Glasgow Norms: Ratings of 5,500 words on nine scales. *Behavior Research Methods*, 1–13.

Sera, M. D., Berge, C. A., & Castillo Pintado, J. del. (1994). Grammatical and conceptual forces in the attribution of gender by English and Spanish speakers. *Cognitive Development*, *9*(3), 261–292.

Silverstein, M. (1981). The limits of awareness. *Sociolinguistic Working Paper Number 84*.

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

Stoet, G., & Geary, D. C. (2018). The gender-equality paradox in science, technology,

engineering, and mathematics education. *Psychological Science*, *29*(4), 581–593.

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