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

this is an abstract

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, non-mutually exclusive source of information is linguistic. 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 gendered contexts more generally.

Past research shows that even young children are sensitive to such linguistically delivered information. 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 an important source of information for forming cultural stereotypes, two people with similar direct experiences, but different linguistic experiences may come to have different stereotypes.

To measure individuals' cultural stereotypes, researchers have commonly used a task that uses reaction time as an index of the degree of psychological association between two concepts (e.g. males-career and women-family), known as the *Implicit Association Task* (IAT; Greenwald, McGhee, & Schwartz, 1998). As it turns out, various biases studied using IATs can be predicted from language using distributional semantics which model word meanings as the contexts in which the word occurs. Caliskan, Bryson, and Narayanan (2017; henceforth *CBN*) measured the distance in vector space between the words presented to participants in the IAT task. CBN found that these distance measures were highly correlated with reaction times in the behavioral IAT task. For example, CBN find a bias to associate males with career and females with family in the career-gender IAT, suggesting that the biases measured by the IAT are also found in the lexical semantics of natural language. CBN only measured semantic biases in English. Here, we extend CBN's method to 26 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 a gender-career IAT previously administered (N = 663,709; Nosek, Banaji, & Greenwald, 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 25% of languages have some type of grammatical gender (Corbett, 1991) 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). But more importantly for present purposes, 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 the degree to which gender in different languages is encoded in their distributional structure. That is, our analysis is not concerned with *what* is being said about typical gender roles, but rather the statistical patterns to which *potentially* gendered words are gendered by different languages. In Study 1b we use this information to predict previously collected responses on an gender-career IAT. In Study 2 we examine whether the extent to which different languages use different forms for occupations (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

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 764,520 participants from 39 countries, with a median of 1,311 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 L2 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.

At the participant level, implicit bias scores were correlated with participant age such that older participants tended to have a larger gender bias than younger participants (r = 0.06, $p < .0001$). Male participants ($M = 0.32$, $SD = 0.39$) had a significantly smaller implicit gender bias than female participants ($M = 0.42$, $SD = 0.36$; $t = 105.60$, $p < .0001$), 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 = -114.08, $p < .0001$).

Because we did not have language information at the participant level in the remaining

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

analyses we examine gender bias and its predictors at the country level. To account for covariates of gender 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 = 0.17). 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 = 0.10).

Does the implicit and explicit biases 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 = -0.59, 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 = 0.08, p = 0.65). 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 (for which participant age was controlled for): 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: Gender bias and semantics

Are participants' implicit and explicit gender biases predictable from biases found in the semantic structure of the dominant language in the country they're in? 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 hypotheses predict a positive correlation between the measured biases and biases present in language.

To model word meanings, we use semantic embeddings derived from trying to predict words from other words in a large corpus. The underlying 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

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.

Methods. 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 gender 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 trained using the fastText algorithm (a variant of word2vec; Bojanowski et al., 2016; Joulin, Grave, Bojanowski, & Mikolov, 2016); The subtitle corpus is a smaller corpus of spoken language, trained using the XX algorithm.

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 and Discussion. 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 ($M = 4.10$; $SD = 0.92$; $r_{\text{subtitles}} = 0.63$; $p < .0001$; $r_{\text{Wikipedia}} = 0.59$; $p < .0001$; 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 participants and compare it with estimates of behavioral gender bias from the Project Implicit data set.

Methods. For each country represented in our analysis of the Project Implicit, we identified the most frequently spoken language in each country using Ethnologue (Simons & (eds.)., 2018). This included a total of 26 unique languages. For each language, we then obtained translations from native speakers for the stimuli 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,"

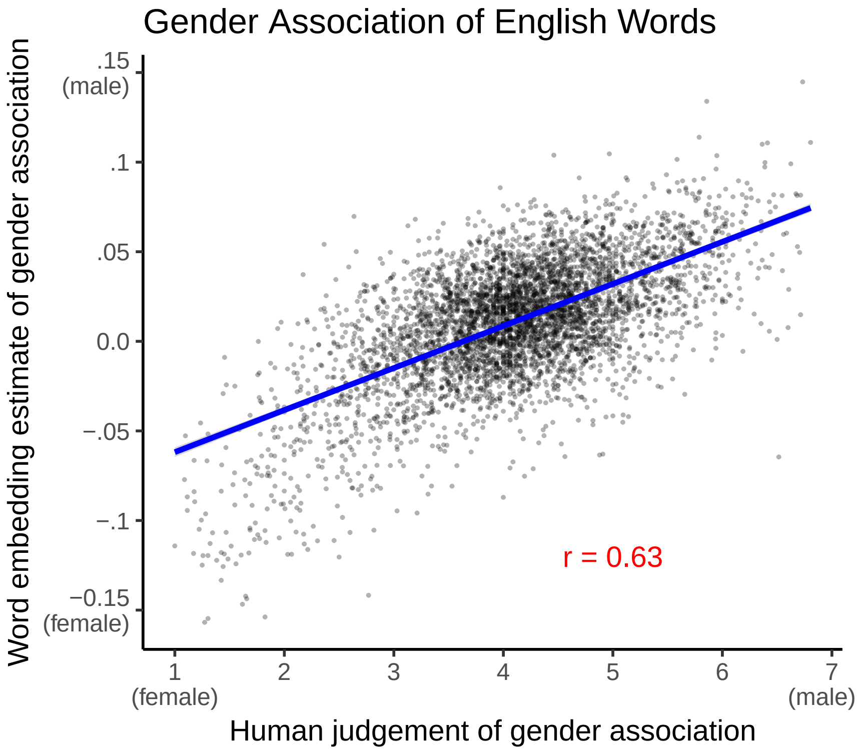

Figure 1. Word estimates of gender bias from the Subtitle-trained embedding model as a function of human judgments of gender bias (Study 1a). Each point corresponds to a word. Larger numbers indicate stronger association with males. Blue line shows linear fit and the error band corresponds to a standard error (too small to be visible).

"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), a set of 8 Attribute Words associated with the concept "career" ("career," "executive," "management," "professional," "corporation," "salary," "office," "business") and 8 Attribute Words associated with the concept "family" ("family," "home," "parents," "children," "cousins," "marriage," "wedding," "relatives"). For one language, Tagalog, we were unable to obtain translations from a native speaker, and so translations were gathered from several translations sources. All analyses remain the same when this language is excluded [check this].

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.

We calculated gender bias estimates from word-embedding models that had been trained on texts Wikipedia (Bojanowski et al., 2016) and subtitles from movies and TV shows (Lison & Tiedemann, 2016, as in Study 1a; Van Paridon & Thompson, in prep.) in each of our target languages. 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 lead 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 9 different language families. We then compared estimates of linguistic gender bias for each language from each model to the behavioral IAT gender bias estimated from Project Implicit, averaging across countries whose participants speak the same language.

As before, we included two country level variables in our analysis, percentage of women in STEM fields and median country age. To obtain language-level estimates of these variables, we took the mean across countries whose participants speak the same primary language.

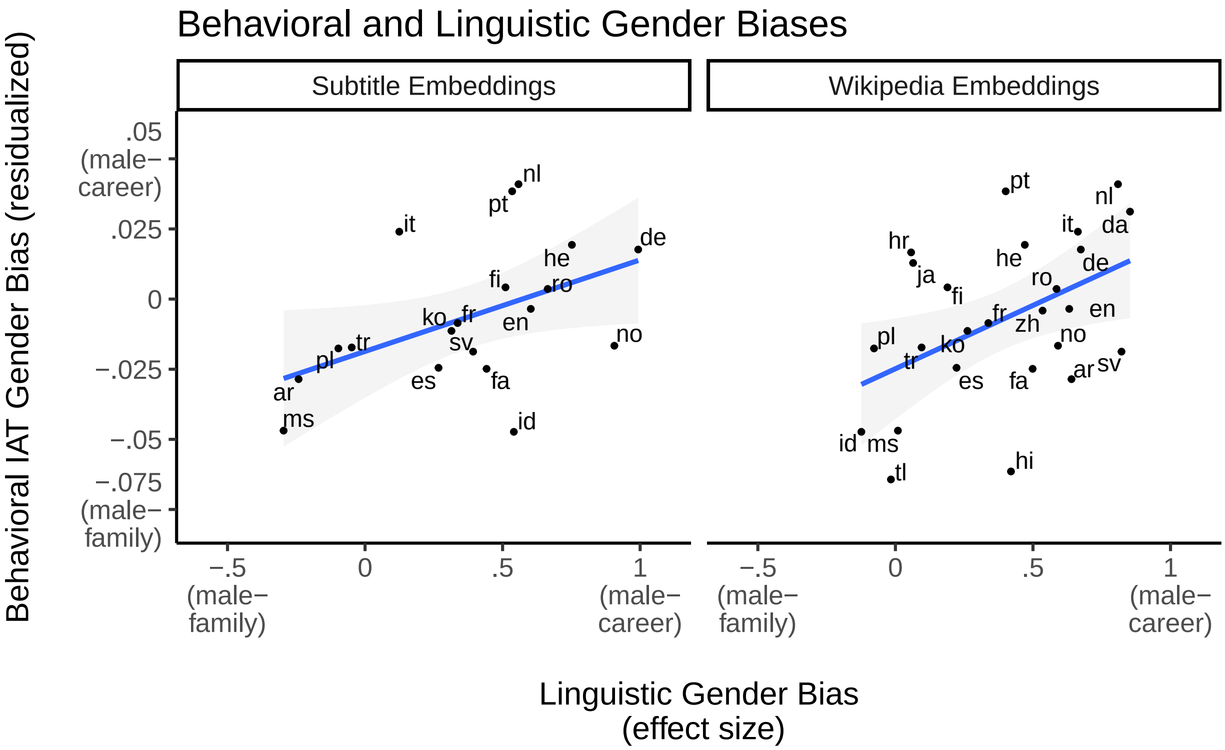

Figure 2. Residualized behavioral IAT gender bias as a function of linguistic gender bias, with each point corresponding to a language (Study 1b). Linguistic biases are estimated from models trained on text in each language from a subtitle corpus (left) and a Wikipedia corpus (right). Larger values indicate a larger bias to associate men with the concept of career and men with the concept of family. Error bands indicate standard error of the model estimate.

Results. There was no overall difference in the estimates of gender bias between models trained on the subtitle corpora versus the Wikipedia corpora ($t(19)$ = -0.06, p = 0.95). We next asked about the relationship between estimates of gender bias for each language and implicit gender bias of participants from countries where that language was dominant (and, we assume, was the native language of most of these individuals). Implicit gender bias were positively correlated with estimates of language bias from both Subtitles (r = 0.54, p = 0.01) and Wikipedia (r = 0.47, p = 0.02; Fig. 2, Table 1 shows the language-level correlations between all variables). The relationship between behavioral bias and language bias remained reliable after partialling out the effect of median country age (Subtitles: r = 0.46, p = 0.02; Wikipedia: r = 0.41, p = 0.04). Linguistic gender bias was not correlated with explicit gender bias (Subtitles: r = -0.06, p = 0.8; Wikipedia: r = 0.31, p = 0.13). Finally, estimates of language bias from the Subtitles corpus were correlated with the objective measure of gender equality, percentage of women in STEM fields (r = -0.55, p = 0.02), though this relationship was not reliable for the Wikipedia corpus (r = -0.19, p = 0.4).

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 from countries in which the language statistics (of the country's 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 these hypotheses by investigating whether the gender-career bias is associated with *structural* features of language that are to a large extent independent of the informational *content* of language.

Study 2: Gender bias and lexicalized gender

If the correlation between language statistics play a causal role in shaping psychological gender biases, we predicted that those statistics would be influenced by the presence of explictly-marked gender distinctions referring to people. Languages make gender distinctions on words referring to people in order to indicate the biological sex of the referent, as in "waiter" versus "waitress" in English. We hypothesized that the presence of these kinds of distinctions might *lead to* more gender biased language statistics for those words. This prediction is a stronger test of the language-as-causal hypothesis because these gender markings are fossilized as part of the lexicon, and thus less likely to be caused by people's gender biases.

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)), or 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.

Thus, in Study 2, we asked whether languages that make gender distinctions more often on words referring to people tend to have more biased language statistics for those words. In this cases, gender marking highlights the gender of the person and thereby might exaggerate the gender biases present in the language. To test this possibility, we used word embedding models to examine the semantic associates that emerge from language statistics for a set words referring to occupations (e.g., "nurse" and "waiter"). We hypothesized that words referring to occupations that contained lexicalized gender distinctions would be more likely to have biased language statistics. To the extent that language statistics are causally related to people's implicit biases, we also hypothesized that speakers of languages with more biased language statistics for these words would also have larger overall psychological gender biases, as measured by the IAT.

Method. We identified 20 words referring to occupations that were relatively common and balanced for their perceptions of gender bias in the workforce, on the basis of previously-collected gender perception norms (Misersky et al., 2014). We then translated these words into each of the 26 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 male and female forms that were identical for each item, and then averaged across items within each language. This measure reflects the degree to which a language marks gender lexically, with larger values indicating more gender-neutral forms. We also estimated the extent to which each occupation word was gender biased in its language statistics using word embedding models trained. This measure was obtained using the same procedure as in Study 1a and 1b based on models trained on both subtitle and Wikipedia corpora in each language. Larger values indicate greater gender bias (larger difference between associations to females vs. males). We compared these measures to the psychological gender measures described in Study 1b (implicit association bias effect size adjusted for age, sex and block order, and explicit bias score).

Results.

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 Behavioral IAT	Residualized Explicit Bias	Percent Women in STEM	Language IAT (Subtitles)	Language IAT (Wikipedia)	Occupation Bias (Subtitles)	Occupation Bias (Wikipedia)	Prop. Gender-Distinct Labels	Median Country Age
Residualized Behavioral IAT		.1 (.23)	.59** (.5*)	.51* (.45*)	.35+ (.22)	.62** (.54**)	.54** (.46*)	.56** (.45*)	.62**
Residualized Explicit Bias	.1 (.23)		.09 (.05)	.03 (.01)	.14 (.2)	.28 (.35+)	.35+ (.41*)	.05 (.11)	.13
Percent Women in STEM	.59** (.5*)	.09 (.05)		.55* (.51**)	.09 (.02)	.39 (.29)	.31 (.22)	.32 (.21)	.37+
Language IAT (Subtitles)	.51* (.45*)	.03 (.01)	.55* (.51**)		.48* (.43*)	.42+ (.36+)	.39+ (.34+)	.26 (.18)	.27
Language IAT (Wikipedia)	.35+ (.22)	.14 (.2)	.09 (.02)	.48* (.43*)		.28 (.19)	.47* (.42*)	.22 (.12)	.3
Occupation Bias (Subtitles)	.62** (.54**)	.28 (.35+)	.39 (.29)	.42+ (.36+)	.28 (.19)		.8** (.77**)	.75** (.71**)	.36
Occupation Bias (Wikipedia)	.54** (.46*)	.35+ (.41*)	.31 (.22)	.39+ (.34+)	.47* (.42*)	.8** (.77**)		.66** (.61**)	.31
Prop. Gender-Distinct Labels	.56** (.45*)	.05 (.11)	.32 (.21)	.26 (.18)	.22 (.12)	.75** (.71**)	.66** (.61**)		.38+
Median Country Age	.62**	.13	.37+	.27	.3	.36	.31	.38+	

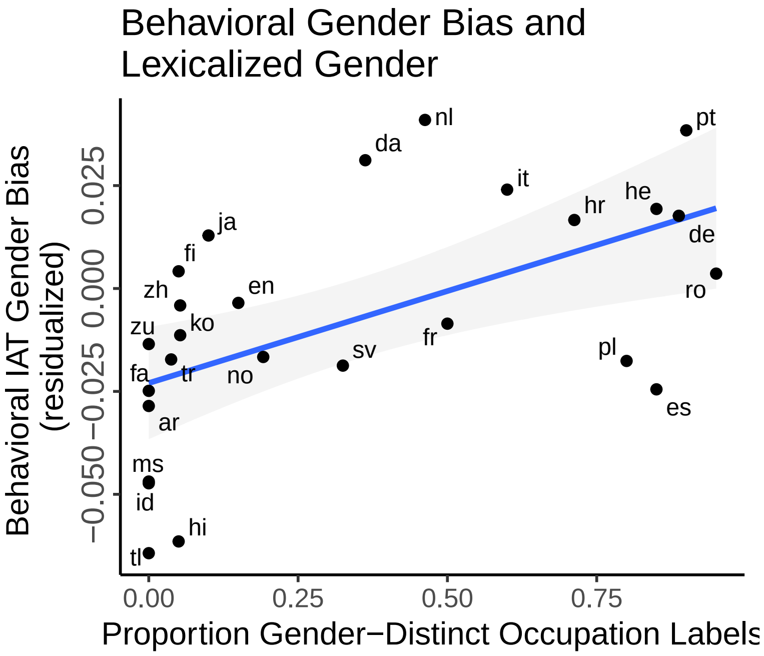

Figure 3. Residualized behavioral IAT gender bias as a function of the proportion of gender-netural labels for set of words referring to occupations. Error bands indicate standard error of the model estimate.

Languages with more distinct forms for male and female referents tended to have speakers with higher psychological gender bias (M = 0.34, SD = 0.36; r = 0.56, p < .01), even after partialling out the effect of median country age (r = 0.45, p = 0.02; Table 1).

We next examined whether having distinct forms for males and females in a particular occupation was associated with greater gender bias in the language statistics. 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 overlap between male and female forms for that word, with random intercepts and slopes by language. Degree of form overlap was a strong predictor of language statistic for models trained on both the subtitle corpus (β = 0.59; SE = 0.07; t = 8.72) and Wikipedia corpus (β = 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: $r = 0.66$, $p < .01$).

Finally, we examined the relationship between gender bias in language statistics and psychological gender bias 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 level of languages between bias in language statistics *for occupation words* and behavioral biases of speakers of that langauge. Consistent with this prediction, gender bias in language statistics for occpuation words was positively correlated with behavioral gender bias for both language models (Subtitle corpus: $r = 0.62$, $p < .01$; Wikipedia corpus: $r = 0.54$, $p = 0.01$), and remained reliable after partialling out the effect of median country age (Subtitle corpus: $r = 0.54$, $p = 0.01$; Wikipedia corpus: $r = 0.46$, $p = 0.02$; Figure 4). To examine the relative predictive power of gender bias in language statistics and proportion form overlap, we fit a linear regression predicting behavioral gender bias with both measures, controlling for median country age. . Bias in language statistics did not reliably predict explicit gender bias. [MEDIATION MODEL]

Discussion

TO SORT OUT:

- Look at native vs. dictionary tranlsated languages
- remove low frequency words - based on google hits?
- deal with missing zh and ja translations for occupations, and others
- try mediation analysis?

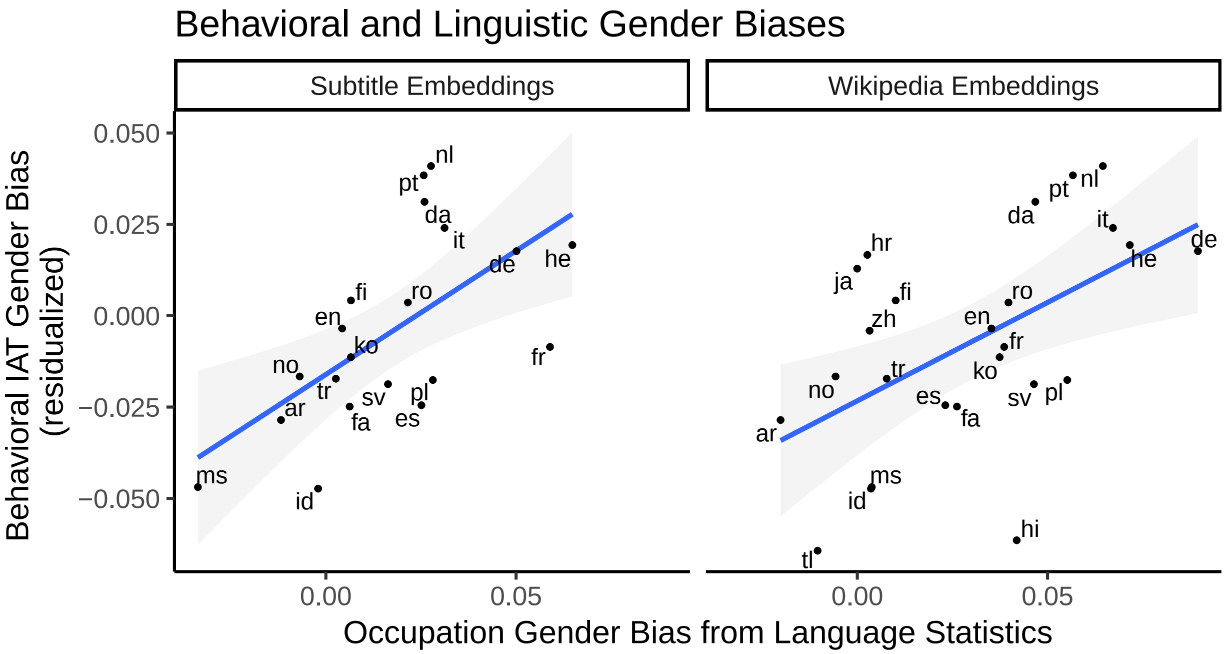

Figure 4. Residualized behavioral IAT gender bias as a function of mean gender bias of words referring to occupations, with each point corresponding to a language (Study 2). Linguistic biases are estimated from models trained on text in each language from a subtitle corpus (left) and a Wikipedia corpus (right).

General Discussion

- IAT is only behavioral in a weak sense
- socialization in development
- implicit vs. explicit difference
- pragmatics related to presense of distinction for kids - must be relevenat! - chesnut stuff
- gender asymmetries in the IAT (women show bigger effect than men?)
- experimental work to get at causality more so

References

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.

Corbett, G. G. (1991). *Gender.* Cambridge: Cambridge University Press.

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.

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 (lrec 2016)*.

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

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.

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.

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.

Simons, G. F., & (eds.)., C. D. F. (Eds.). (2018). Ethnologue: Languages of the world.

Ethnologue: Languages of the World.

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.