Gender Bias in Italian Word Embeddings

Davide Biasion, Alessandro Fabris, Gianmaria Silvello, Gian Antonio Susto
Università degli Studi di Padova
{biasiondav,fabrisal,silvello,sustogia}@dei.unipd.it

Abstract

In this work we study gender bias in Italian word embeddings (WEs), evaluating whether they encode gender stereotypes studied in social psychology or present in the labor market. We find strong associations with gender in job-related WEs. Weaker gender stereotypes are present in other domains where grammatical gender plays a significant role.

1 Introduction

In the literature, the study of gender bias in word embeddings (WEs) is of interest for two main reasons: (i) WEs, as components of automatic decision systems (e.g. job search tools), may contribute to harm some user groups (De-Arteaga et al., 2019); (ii) WEs can be employed as a tool to measure the biases of text corpora (Garg et al., 2018) and systems for automatic text classification or information retrieval (Fabris et al., 2020). In both applications, it is important to isolate the gender-related information in a subspace (Bolukbasi et al., 2016) and subsequently (i) eliminate it via orthogonal projection or (ii) exploit it as a lens to study association of concepts with gender.

A common taxonomy of bias in algorithms concentrates on the types of harm that they may cause (Barocas et al., 2017). **Allocational harms** happen when a limited resource (e.g. jobs) is assigned unfairly to subgroups of a population (e.g. women and men). **Representational harms** arise when groups or individuals are unable to determine their image, which is presented unfavourably or neglected. Autocomplete suggestions in search engines (Noble, 2018; Olteanu et al., 2020) are a clear example of this situation. Query completion suggestions for "why are italian ..." associate diverse concepts to the country and its inhabitants. Italians contribute very little to these results as they are unlikely to search information about themselves in English.

Italian WEs have been developed (Berardi et al., 2015; Bojanowski et al., 2017) and analyzed (Tripodi and Li Pira, 2017), following seminal work in English; analysis of gender bias has unfortunately lagged behind. Our main contribution is to close this gap, by undertaking a systematic study of gender stereotypes in Italian WEs, adapting established approaches that assess gender bias in English WEs.

2 Related work

Gender stereotypes are representational harms which influence the lives of women and men both descriptively and prescriptively, shaping the qualities, priorities and needs that members of each gender are expected to possess (Ellemers, 2018). In seminal work, Bolukbasi et al. (2016) uncover problematic associations with gender in English WEs. Their approach to identify gender information is adapted to Italian in Section 3.1.1. Caliskan et al. (2017) study the stereotypical association of gender with dichotomies such as career and family, science and arts, following the Implicit Association Test (IAT - Greenwald et al. (1998)).We recall their approach in Section 3.1.2. WEs of jobs have also been analyzed extensively due to their potential for allocational harms in resume search engines (De-Arteaga et al., 2019; Prost et al., 2019) and representational harms (Caliskan et al., 2017), e.g. in general purpose search engines (Kay et al., 2015). Grammatical gender has been found to interact strongly with semantic gender in Spanish and German (McCurdy and Serbetci, 2020), showing that the study of bias in gendered languages poses an additional challenge. We adapt

these experiments to the Italian language, detailing our approach in Sections 3-5.

3 Gender in Italian WEs

3.1 Identifying gender information

3.1.1 Gender score

To identify a vectorial subspace which encodes information about gender, we follow Bolukbasi et al. (2016) by building a list of gender definitional pairs: [*lui* (*he*), *lei* (*she*)], [*uomo* (*man*), *donna* (*woman*)], [*padre* (*father*), *madre* (*mother*)], [*marito* (*husband*), *moglie* (*wife*)], [*fratello* (*brother*), *sorella* (*sister*)], [*maschio* (*male*), *femmina* (*female*)].

These pairs are built so that the second word denotes a female entity and the first word is, semantically, its male counterpart. Moreover, given we are interested in capturing *semantic* information about gender, while avoiding entanglement with *grammatical* gender, we ensure that the words in a pair do not derive from the same root via inflection. An example of pair discarded due to this criterion is [*figlio* (*son*), *figlia* (*daughter*)].

Principal Component Analysis. We perform a Principal Component Analysis (PCA) on the six vector differences resulting from each gender definitional pair. The first eigenvalue dominates the remaining ones, with the first PC explaining 57% of variance. We normalize the first PC and consider it the main *gender direction*, denoted by $\mathbf{g}_{\text{PCA}}$.

This is an established procedure to isolate the direction that captures most of the information about gender (Bolukbasi et al., 2016; Ethayarajh et al., 2019). In other words, by finding the direction that best fits the six vector differences $(\overrightarrow{lui} - \overrightarrow{lei}, \overrightarrow{uomo} - \overrightarrow{donna}, \dots)$, we aim to obtain a direction that summarizes them.

Vector differences. To evaluate the robustness of this approach and highlight potential anomalies, we also consider each vector difference on its own, defining six unit length *gender directions* $\mathbf{g}_{\text{diff}_i}$:

$$\mathbf{g}_{\text{diff}_0} = \overrightarrow{lui} - \overrightarrow{lei} \qquad \mathbf{g}_{\text{diff}_3} = \overrightarrow{marito} - \overrightarrow{moglie}$$
$$\mathbf{g}_{\text{diff}_1} = \overrightarrow{uomo} - \overrightarrow{donna} \qquad \mathbf{g}_{\text{diff}_4} = \overrightarrow{fratello} - \overrightarrow{sorella}$$
$$\mathbf{g}_{\text{diff}_2} = \overrightarrow{padre} - \overrightarrow{madre} \qquad \mathbf{g}_{\text{diff}_5} = \overrightarrow{maschio} - \overrightarrow{femmina}$$

Gender score computation. Given a word w, let us indicate with $\mathbf{w}$ its corresponding word vector. Let us consider any of the gender directions $\mathbf{g}$ defined above. We call *gender score* the normalized projection of $\mathbf{w}$ onto the direction $\mathbf{g}$, defined as

$$s_{\mathbf{g}}(w) = \mathbf{w} \cdot \mathbf{g}/(|\mathbf{w}||\mathbf{g}|). \tag{1}$$

This scalar captures associations of w along gendered lines. Informally, a highly positive value means that $\mathbf{w}$ is closer to the male terms of the pairs than to the female ones, while a strongly negative value entails the opposite.

3.1.2 WEAT

The Implicit Association Test (IAT - Greenwald et al. (1998)) is an assessment developed in cognitive psychology to measure subconscious associations between categories and concepts. It is commonly employed to assess implicit stereotypes in people. The *Word Embedding Association Test* (WEAT - Caliskan et al. (2017)) is a technique inspired by the IAT to measure associations between concepts in WEs. Let X and Y be two equal-sized sets of target words and A and B two sets of attribute words, e.g., $X = \{\textit{programmer}, \textit{engineer}\}$, $Y = \{\textit{nurse}, \textit{teacher}\}$, $A = \{\textit{man}, \textit{male}\}$, $B = \{\textit{woman}, \textit{female}\}$. Let $\cos(\mathbf{a}, \mathbf{b})$ be the cosine similarity between the word vectors $\mathbf{a}$ and $\mathbf{b}$. The differential association of a word w (taken from X or Y) with the attribute sets A and B is measured as

$$c(w, A, B) = \text{mean}_{a \in A} \cos(\mathbf{w}, \mathbf{a}) - \text{mean}_{b \in B} \cos(\mathbf{w}, \mathbf{b}). \tag{2}$$

The normalized differential association between targets and attributes is defined as

$$d = \frac{\text{mean}_{x \in X} c(x, A, B) - \text{mean}_{y \in Y} c(y, A, B)}{\text{std-dev}_{w \in X \cup Y} c(w, A, B)}. \tag{3}$$

This is called effect size in statistics, and summarizes how different the quantity $c(w, A, B)$ is, when evaluated on elements of target set X as opposed to target set Y. It is computed as a difference of means within each set, divided by overall standard deviation.

Gender score and WEAT. It is worth noting that, when $|A| = |B| = 1$, WEAT is almost equivalent to the gender score defined in Section 3.1.1. Let $A = \{a_0\}$ and $B = \{b_0\}$ be the sets of attribute words. Since we are using normalized vectors and the distributive property holds for the dot product, then

$$\begin{aligned} c(w, A, B) &= \cos(\mathbf{w}, \mathbf{a}_0) - \cos(\mathbf{w}, \mathbf{b}_0) \\ &= \mathbf{w} \cdot (\mathbf{a}_0 - \mathbf{b}_0) = \mathbf{w} \cdot \mathbf{g} = s_{\mathbf{g}}(w). \end{aligned} \tag{4}$$

3.2 Handling grammatical gender

Italian is a gendered language, wherein grammatical gender is assigned to all nouns. Within a sentence, each word is surrounded by other words of agreeing grammatical gender. This phenomenon, called *grammatical gender agreement*, in conjunction with the *distributional hypothesis* (Harris, 1954), plays an important role when training WEs. Due to these properties, words that share the same grammatical gender to have similar vector representations. Accordingly, grammatical and semantic gender become entangled in WEs (McCurdy and Serbetci, 2020; Gonen et al., 2019). As a consequence, when computing the gender score, we tend to obtain positive values for (grammatically) masculine terms and a negative score for feminine ones, making stereotypical association more noisy and harder to study.

Mean gender score. To compute the gender score (Equation 1) for gendered words that have both a feminine and a masculine version, we propose the following approach. Let us indicate with w_f and w_m the feminine and masculine version of a gendered word w. We define their gender score as

$$s_{\text{mean}_\mathbf{g}}(w) = (s_\mathbf{g}(w_f) + s_\mathbf{g}(w_m))/2. \quad (5)$$

Averaging the masculine and feminine version with equal weights corresponds to giving both versions of the word the same importance. Different approaches, based for instance on word frequency, may be applicable in other contexts.

Orthogonal projection. Some nouns cannot be inflected into the opposite grammatical gender, making the above approach impractical. An example is *ufficio* (*office*). In this context, we propose to mitigate the effect of grammatical gender by re-embedding every word through an orthogonal projection. We build a list of 138 inflected word pairs. Each pair consists of the feminine and masculine inflections of the same root, such as *cara* and *caro* (*dear*), which only differ in grammatical gender. We take the embedding of both words in a pair and compute their difference.

We perform PCA on these vector differences. The resulting PCs span a subspace U that contains most of the variance due to grammatical gender. To reduce the influence of grammatical gender, we re-embed vectors by projecting them on the orthogonal complement of U. In other words, given a word embedding $\mathbf{w}$, let us call $\texttt{proj}_U\mathbf{w}$ its orthogonal projection onto the "grammatical gender subspace" U. We propose re-emdedding every word vector $\mathbf{w}$ to

$$\mathbf{w}' = \frac{w - \texttt{proj}_U\mathbf{w}}{|w - \texttt{proj}_U\mathbf{w}|}. \quad (6)$$

By means of this procedure, we obtain a new set of WEs. By construction, in this new embedding space, grammatical gender should have a lower influence on the geometry of word vectors.

4 Datasets and embeddings

To study gender bias we use WEs trained on two different datasets for the Italian language, both made available by FastText (Bojanowski et al., 2017; Grave et al., 2018). The first group of vector representations, which we refer to as `wiki`, consists of word vectors trained on a 2016 Wikipedia dump (Bojanowski et al., 2017).[1] The second group of word vectors (labeled `wiki-cc`) was trained on the May 2017 Common Crawl[2] and the Wikipedia dump from September 11, 2017 (Grave et al., 2018).

We compare our results from the analyses on Italian WEs with results on their English counterpart. To this end, we also download two sets of FastText WEs trained on the English version of the same corpora, i.e. the English counterparts of `wiki` and `wiki-cc`. Given Wikipedia is a more curated source, we expect to find weaker stereotypes in `wiki` than in `wiki-cc` for both languages. As a pre-processing step we normalize every word vector to unit length.

Census data about the labor market is required to analyse the correlation between the gender gap in professions and the gender score of the respective WEs. The statistics on the American occupation and gender representation are readily available (Census Bureau, 2019). For their Italian counterpart, we retrieve statistics about occupation participation from several institutions, including professional chambers (Comitato Unitario Permanente degli Ordini e Collegi Professionali, Confprofessioni) and academic databases (AlmaLaurea).[3]

Finally, in order to perform the Word Embedding Association Tests (WEAT), we need sets of

[1] The authors provide no information about which Wikipedia dump they use.

[2] Common Crawl is a corpus of web pages, aimed at representing "a copy of the internet" at a given time. The authors train WEs on pages written in Italian, exploiting language identification as preliminary step for their pipeline.

[3] The detailed list of sources is available upon request.

target and attribute words in Italian. The sets of target words for the gender-science WEAT (Section 5.2) are derived from the Italian version of IAT;[4] those for the gender-career WEAT (Section 5.3) were unavailable and have been translated by the authors of this work from the original IAT (Greenwald et al., 1998).

5 Experiments

5.1 Occupations

This experiment investigates gender representation for different jobs in Italy and their association with gender-realted information in WEs, following studies on the English language (De-Arteaga et al., 2019; Garg et al., 2018; Prost et al., 2019). For each occupation, we compute its gender score using the different gender directions defined in Section 3.1.1, namely $\mathbf{g}_{\text{PCA}}$ and $\mathbf{g}_{\text{diff}_i}$, $i \in \{0 \dots 5\}$. We calculate the plain gender score for the ungendered occupations (Equation 1) and the mean gender score for occupations characterized by grammatical gender (Equation 5).

We compute Pearson's correlation r between the gender scores and the percentage of women employed in each profession. The same analyses are carried out on English WEs, restricting them to the same set of occupations considered in Italian. Results are summarized in Table 1 and Figure 1, showing that Italian WEs consistently capture information about different gender representation in jobs. Informally, this mean that ordering jobs by percentage of women and by projection on a gender direction yields similar results. The right pane of Figure 1 demonstrates the significant effect of grammatical gender.

5.2 Science and Arts

In this WEAT, the sets of target words for Science and Arts, taken from the Italian version of the IAT, are: $X = \{$*biologia* (*biology*), *fisica* (*physics*), *chimica* (*chemistry*), *matematica* (*mathematics*), *geologia* (*geology*), *astronomia* (*astronomy*), *ingegneria* (*engineering*)$\}$, $Y = \{$*filosofia* (*philosophy*), *umanesimo* (*humanism*), *arte* (*arts*), *letteratura* (*literature*), *italiano* (*italian*), *musica* (*music*), *storia* (*history*)$\}$. The sets of male and female attributes are taken from the gender definitional pairs (Section 3.1.1): $A = \{$*lui*, *uomo*, *padre*, *marito*, *fratello*, *maschio*$\}$, $B = \{$*lei*, *donna*, *madre*, *moglie*, *sorella*, *femmina*$\}$. We compute the effect size d and the p-value using the whole attribute sets A and B, and label this analysis "all". Moreover, we also perform the WEAT test over single word pairs, e.g. $A = \{$*lui*$\}$, $B = \{$*lei*$\}$. Results are reported in Table 2. We find no stereotypical association in the expected direction. We hypothesize that this is due to the feminine grammatical gender of all science-related target words, deferring a more detailed analysis to Section 5.4.

[4] https://implicit.harvard.edu/implicit/italy/takeatest.html

	wiki-cc	wiki
IT	r (p)	r (p)
$\mathbf{g}_{\text{PCA}}$	-0.634 (1.3×10^{-4})***	-0.589 (4.9×10^{-4})***
$\mathbf{g}_{\text{diff}_0}$	-0.664 (4.7×10^{-5})***	-0.490 (5.1×10^{-3})***
$\mathbf{g}_{\text{diff}_1}$	-0.594 (4.3×10^{-4})***	-0.528 (2.3×10^{-3})***
$\mathbf{g}_{\text{diff}_2}$	-0.575 (7.1×10^{-4})***	-0.537 (1.8×10^{-3})***
$\mathbf{g}_{\text{diff}_3}$	-0.401 (2.5×10^{-2})**	-0.160 (3.9×10^{-1})
$\mathbf{g}_{\text{diff}_4}$	-0.658 (5.7×10^{-5})***	-0.599 (3.8×10^{-4})***
$\mathbf{g}_{\text{diff}_5}$	-0.358 (4.8×10^{-2})**	-0.205 (2.7×10^{-1})
EN	r (p)	r (p)
$\mathbf{g}_{\text{PCA}}$	-0.830 (2.0×10^{-6})***	-0.707 (2.3×10^{-4})***

Table 1: Results of the Occupation analysis. Statistical significance is marked as * for $p < 0.1$, ** for $p < 0.05$ and *** for $p < 0.01$.

	wiki-cc	wiki
IT	d (p)	d (p)
all	-0.172 (6.3×10^{-1})	-0.140 (5.9×10^{-1})
$\mathbf{g}_{\text{diff}_0}$	-0.464 (7.9×10^{-1})	-0.396 (7.5×10^{-1})
$\mathbf{g}_{\text{diff}_1}$	-0.016 (5.1×10^{-1})	-0.064 (5.4×10^{-1})
$\mathbf{g}_{\text{diff}_2}$	-0.408 (7.5×10^{-1})	-0.152 (6.1×10^{-1})
$\mathbf{g}_{\text{diff}_3}$	-0.002 (5.0×10^{-1})	0.271 (3.3×10^{-1})
$\mathbf{g}_{\text{diff}_4}$	-0.127 (6.0×10^{-1})	-0.174 (6.2×10^{-1})
$\mathbf{g}_{\text{diff}_5}$	-0.144 (6.1×10^{-1})	-0.195 (6.3×10^{-1})
EN	d (p)	d (p)
all	1.420 (1.5×10^{-3})***	1.304 (3.2×10^{-3})***

Table 2: Results of the Science and Arts WEAT. Statistical significance is marked as * for $p < 0.1$, ** for $p < 0.05$ and *** for $p < 0.01$.

5.3 Career and Family

In essence, the Career and Family WEAT is very similar to the Science and Arts WEAT; the only difference is in the sets of target words. The target sets are translated into Italian from the original English IAT as follows: $X = \{$*esecutivo* (*executive*), *management* (*management*), *professionale* (*professional*), *azienda* (*corporation*), *stipendio* (*salary*), *ufficio* (*office*)$\}$, $Y = \{$*casa* (*home*),

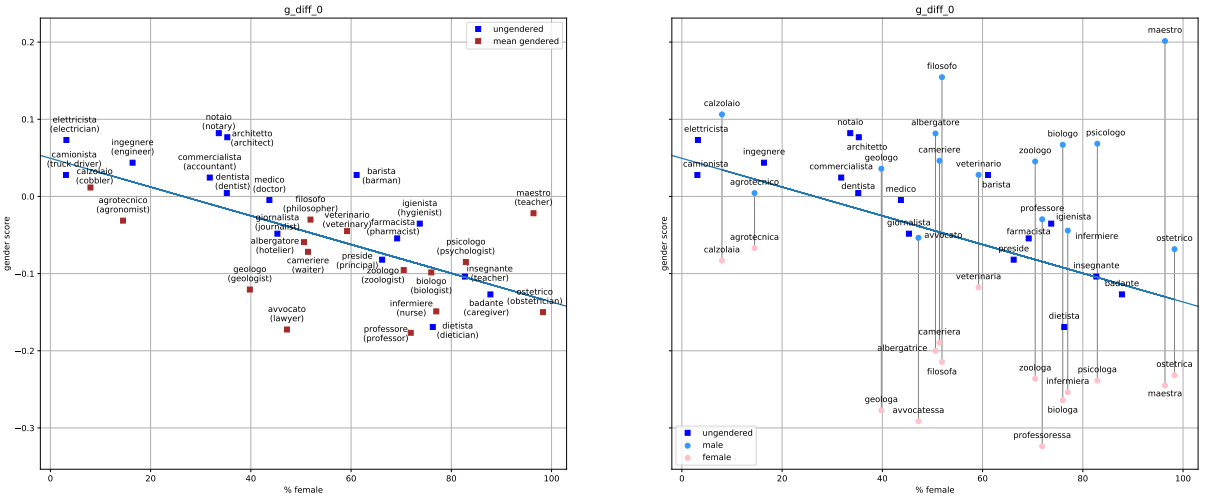

Figure 1: (Left) Gender score of occupations $\mathbf{g}_{\text{diff}_0}$, on the y axis, vs percentage of women in that occupation, represented on the x axis. (Right) Same variables, including both feminine and masculine versions of gendered jobs. Translations, omitted for readability, can be found in the left pane.

genitori (*parents*), *bambini* (*children*), *famiglia* (*family*), *cugini* (*cousins*), *matrimonio* (*marriage*), *nozze* (*wedding*), *parenti* (*relatives*)}. Results are summarized in Table 3. Stereotypical associations for `wiki-cc` WEs are present but weak, whereas they are more significant for `wiki`.

	`wiki-cc`	`wiki`
IT	d (p)	d (p)
all	0.838 (5.3×10^{-2})*	1.351 (2.7×10^{-3})***
$\mathbf{g}_{\text{diff}_0}$	0.457 (2.0×10^{-1})	1.172 (8.7×10^{-3})***
$\mathbf{g}_{\text{diff}_1}$	1.265 (4.7×10^{-3})***	1.512 (5.4×10^{-4})***
$\mathbf{g}_{\text{diff}_2}$	0.614 (1.2×10^{-1})	1.181 (8.3×10^{-3})***
$\mathbf{g}_{\text{diff}_3}$	0.299 (2.9×10^{-1})	0.876 (4.4×10^{-2})**
$\mathbf{g}_{\text{diff}_4}$	0.952 (3.2×10^{-2})**	0.898 (4.2×10^{-2})**
$\mathbf{g}_{\text{diff}_5}$	0.713 (8.7×10^{-2})*	-0.566 (8.6×10^{-1})
EN	d (p)	d (p)
all	1.879 (0.0×10^{0})***	1.568 (2.3×10^{-4})***

Table 3: Results of the Career and Family WEAT. Statistical significance is marked as * for $p < 0.1$, ** for $p < 0.05$ and *** for $p < 0.01$.

5.4 Mitigating the effect of grammatical gender

In this section we quantify the extent to which the semantic gender information (Section 3.1) is influenced by grammatical gender, and test one approach designed to mitigate its influence (Section 3.2).

The dataset used in the experiment about job-related WEs (Section 5.1) is suitable for this analysis, as it consists of words which have (i) a *semantic* association with gender, as measured objectively by the percentage of women in each profession and (ii) a *grammatical* association with gender, as half of those words admit a feminine and a masculine version.

We measure the relative strength of semantic and grammatical associations in the proposed gender directions as follows. Let us denote by $\mathcal{S}_g$, the set of job-related words which admit a feminine and a masculine version and by $\Delta_w = s_{\mathbf{g}}(w_m) - s_{\mathbf{g}}(w_f)$ the difference in their gender scores.[5] We compute the average influence of grammatical gender on direction $\mathbf{g}$ (based on set $\mathcal{S}_g$) as

$$\Delta_g = \frac{1}{|\mathcal{S}_g|} \sum_{w \in \mathcal{S}_g} \Delta_w. \tag{7}$$

Visually, this corresponds to the average (signed) length of the vertical lines, in the right pane of Figure 1, connecting the feminine and masculine version of a job-related word.

Moreover, let us denote by $\mathcal{S}$ the complete set of job-related words w_j and by x_j the percentage of women in job w_j. Let us indicate with $s_{\mathbf{g}}(w_j)$ the respective gender score, computed according to Equation 1 or 5, depending on whether w_j admits different masculine and feminine inflections. We define $\max_{\mathcal{S}}(x)$ ($\min_{\mathcal{S}}(x)$) as the maximum (minimum) percentage of women in a job from set

[5] For the sake of brevity, we concentrate on $\mathbf{g}_{\text{PCA}}$; the remaining gender directions ($\mathbf{g}_{\text{diff}_i}$) yield similar results.

$\mathcal{S}$. Furthermore, let us call m the angular coefficient computed by (linearly) regressing $s_{\mathbf{g}}(w_j)$ onto x_j over set $\mathcal{S}$. We compute the full-scale influence of semantic gender on direction $\mathbf{g}$ (based on set $\mathcal{S}$) as

$$\Delta_s = |m(\max_{\mathcal{S}}(x) - \min_{\mathcal{S}}(x))|. \quad (8)$$

Visually, this corresponds to the vertical component of the blue regression line in Figure 1, clipped between $\min_{\mathcal{S}}(x)$ and $\max_{\mathcal{S}}(x)$.

Finally, we compute the relative strength of semantic and grammatical associations in the proposed gender direction as the ratio

$$k = \frac{\Delta_g}{\Delta_s}. \quad (9)$$

The first three rows of Table 4 report Δ_g, Δ_s and k for `wiki-cc` (first column) on the job dataset described in Section 4. The second column concentrates on a set of word embeddings derived from `wiki-cc` by removing information about grammatical gender from every word, via Equation 6.[6] We label this new set of word embbeddings `wiki-cc`$^{\perp}$. In going from `wiki-cc` to `wiki-cc`$^{\perp}$, Δ_g is reduced by over 40% while Δ_s decreases by less than 10%. This indicates that the orthogonal projection procedure reduces the influence of grammatical gender while retaining semantic information which is present in the original version of the WEs, hence the value of k decreases.

The final three rows of Table 4 report summary statistics for stereotypical associations described in Sections 5.1-5.3. Interestingly, the significance of each association is larger for `wiki-cc`$^{\perp}$ than for `wiki-cc`. In particular, the effect size for the Science-Arts WEAT becomes positive, in accordance with the stereotype. We interpret these results as evidence for the hypothesis that grammatical gender confounds and outweighs stereotypical associations in Italian WEs, in line with prior work on gendered languages (McCurdy and Serbetci, 2020).

6 Discussion

We successfully replicated prior analyses about gender-stereotypical associations in English WEs, finding them to be consistently stronger when computed on WEs trained on a weakly curated corpus. To the best of our knowledge, this is a novel result.

For Italian WEs, the picture is more nuanced and tied to grammatical gender. WEs for occupations, which are ungendered or admit a dual form, are robustly associated with gender along a stereotypical direction. Compared against the other stereotypes analysed in this work, this is the strongest association, confirming results from prior work on English WEs (Fabris et al., 2020). In the Science-Arts WEAT, science-related words are all feminine nouns, departing from the expected stereotypical association. Semantic associations with gender are outweighed by grammatical gender in this WEAT, in accordance with prior work on gendered languages (McCurdy and Serbetci, 2020). Our analysis in Section 5.4 demonstrates the importance of grammatical gender in Italian. On the other hand, the Career-Family WEAT features a more balanced distribution of grammatical gender, resulting in a differential association which is in line with gender stereotypes, especially for `wiki`, less so for `wiki-cc`.

In Italian WEs, we find that `wiki` embeddings contain stronger stereotypical associations than `wiki-cc` embeddings for the Career-Family WEAT. This disconfirms our hypothesis that WEs trained on a less curated corpus (`wiki-cc`) would encode stereotypes more strongly. Finally, we find no consistent property connected to specific gender directions $\mathbf{g}_{\text{diff}_i}$. Across different corpora and stereotypes, the aggregated analyses (la-

	`wiki-cc`	`wiki-cc`$^{\perp}$
Occupations ($\mathbf{g}_{\text{PCA}}$)		
Δ_g	0.41	0.22
Δ_s	0.23	0.20
k	1.79	1.09
r (p)	-0.63 (1.3×10^{-4})***	-0.68 (2.2×10^{-5})***
Science & Arts (all)		
d (p)	-0.17 (6.3×10^{-1})	0.73 (9.7×10^{-2})*
Career & Family (all)		
d (p)	0.84 (5.3×10^{-2})*	1.21 (6.1×10^{-3})***

Table 4: Importance of semantic and grammatical gender before (`wiki-cc`) and after (`wiki-cc`$^{\perp}$) projecting WEs onto the orthogonal complement of the grammatical gender subspace (Equation 6). Where applicable, statistical significance is marked as * for $p < 0.1$, ** for $p < 0.05$ and *** for $p < 0.01$.

[6] In this experiment, the grammatical gender subspace U is spanned by the first PC.

belled "all" and $\mathbf{g}_{\text{PCA}}$) provide a reasonable summary of the stereotypical associations encoded in the single gender directions $\mathbf{g}_{\text{diff}_i}$.

7 Conlusion

Overall, we have analyzed gender bias in Italian WEs, adapting existing techniques and gathering data where required. We looked for stereotypical associations with gender-imbalanced professions, Career and Family, Science and Arts, finding significant associations in 2 out of 3. As expected from prior work (Gonen et al., 2019; McCurdy and Serbetci, 2020), grammatical gender is a strong confounder in these analyses.

We draw the following preliminary conclusions: (i) Italian WEs seem to have less potential than their English counterparts to systematically reinforce the tested gender stereotypes, mostly due to grammatical gender. However, (ii) the influence of grammatical gender on WEs may cause different harms. As an example, in the context of job search, masculine is likely to be the default choice for queries of recruiters (*male as norm* - e.g. "psicologo" [*psychologist*]). Those queries would likely be closer to male candidates' CVs than equivalent female ones, in some embedded text representations, potentially putting women at a systematic disadvantage. Both points above require further analysis of text retrieval/classification systems based on Italian WEs. Finally, (iii) isolating stereotypical concepts and gendered associations in Italian WEs along a single direction is challenging. The tested WEs show little promise as a reliable measurement tool for gender-stereotypical associations, unless combined with approaches to mitigate the influence of grammatical gender.

References

Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. 2017. The problem with bias: Allocative versus representational harms in machine learning. In *9th Annual Conference of the Special Interest Group for Computing, Information and Society*.

Giacomo Berardi, Andrea Esuli, and Diego Marcheggiani. 2015. Word embeddings go to italy: A comparison of models and training datasets. In *IIR*.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam T. Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, pages 4349–4357.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Census Bureau. 2019. Current population survey. Accessed = 2020-02-12.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios. *Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19*.

Naomi Ellemers. 2018. Gender stereotypes. *Annual Review of Psychology*, 69(1):275–298.

Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. 2019. Understanding undesirable word embedding associations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1696–1705, Florence, Italy, July. Association for Computational Linguistics.

Alessandro Fabris, Alberto Purpura, Gianmaria Silvello, and Gian Antonio Susto. 2020. Gender stereotype reinforcement: Measuring the gender bias conveyed by ranking algorithms. *Information Processing & Management*, 57(6):102377.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Hila Gonen, Yova Kementchedjhieva, and Yoav Goldberg. 2019. How does grammatical gender affect noun representations in gender-marking languages? In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 463–471.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Anthony G. Greenwald, Debbie E. McGhee, and Jordan L. K. Schwartz. 1998. Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, 74(6):1464–1480.

Zellig S Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.

Matthew Kay, Cynthia Matuszek, and Sean A Munson. 2015. Unequal representation and gender stereotypes in image search results for occupations. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 3819–3828. ACM.

Katherine McCurdy and Oguz Serbetci. 2020. Grammatical gender associations outweigh topical gender bias in crosslinguistic word embeddings. *arXiv preprint arXiv:2005.08864*.

Safiya Umoja Noble. 2018. *Algorithms of oppression: How search engines reinforce racism*. nyu Press.

Alexandra Olteanu, Fernando Diaz, and Gabriella Kazai. 2020. When are search completion suggestions problematic? In *Computer Supported Collaborative Work and Social Computing (CSCW)*. ACM, August. Pre-print, paper accepted to CSCW'20.

Flavien Prost, Nithum Thain, and Tolga Bolukbasi. 2019. Debiasing embeddings for reduced gender bias in text classification. *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*.

Rocco Tripodi and Stefano Li Pira. 2017. Analysis of italian word embeddings. In *CLiC-it*.