Unsupervised Discovery of Gendered Language through Latent-Variable Modeling

Hoyle, Alexander Miserlis; Wolf-sonkin, Lawrence; Wallach, Hanna; Augenstein, Isabelle; Cotterell, Ryan

Published in:
Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics

DOI:
10.18653/v1/P19-1167

Publication date:
2019

Document version
Publisher's PDF, also known as Version of record

Document license:
CC BY

Citation for published version (APA):
Unsupervised Discovery of Gendered Language through Latent-Variable Modeling

Alexander Hoyle\textsuperscript{3}  Lawrence Wolf-Sonkin\textsuperscript{f}  Hanna Wallach\textsuperscript{5}

Isabelle Augenstein\textsuperscript{7}  Ryan Cotterell\textsuperscript{f}

\textsuperscript{3}University College London, London, UK
\textsuperscript{f}Department of Computer Science, Johns Hopkins University, Baltimore, USA
\textsuperscript{5}Microsoft Research, New York City, USA
\textsuperscript{7}Department of Computer Science, University of Copenhagen, Copenhagen, Denmark

alexander.hoyle.17@ucl.ac.uk, lawrencews@jhu.edu
hanna@dirichlet.net, augenstein@di.ku.dk, rdc42@cam.ac.uk

Abstract

Studying the ways in which language is gendered has long been an area of interest in sociolinguistics. Studies have explored, for example, the speech of male and female characters in film and the language used to describe male and female politicians. In this paper, we aim not to merely study this phenomenon qualitatively, but instead to quantify the degree to which the language used to describe men and women is different and, moreover, different in a positive or negative way. To that end, we introduce a generative latent-variable model that jointly represents adjective (or verb) choice, with its sentiment, given the natural gender of a head (or dependent) noun. We find that there are significant differences between descriptions of male and female nouns and that these differences align with common gender stereotypes: Positive adjectives used to describe women are more often related to their bodies than adjectives used to describe men.

1 Introduction

Word choice is strongly influenced by gender—both that of the speaker and that of the referent (Lakoff, 1973). Even within 24 hours of birth, parents describe their daughters as beautiful, pretty, and cute far more often than their sons (Rubin et al., 1974). To date, much of the research in sociolinguistics on gendered language has focused on laboratory studies and smaller corpora (McKee and Sherriffs, 1957; Williams and Bennett, 1975; Baker, 2005); however, more recent work has begun to focus on larger-scale datasets (Pearce, 2008; Caldas-Coulthard and Moon, 2010; Baker, 2014; Norberg, 2016). These studies compare the adjectives (or verbs) that modify each noun in a particular gendered pair of nouns, such as boy–girl, aggregated across a given corpus. We extend this line of work by instead focusing on multiple noun pairs simultaneously, modeling how the choice of adjective (or verb) depends on the natural gender\textsuperscript{1} of the head noun's, with sentiment, used to describe men and women, as represented by our model. Colors indicate the most common sense of each adjective from Tsvetkov et al. (2014); black indicates out of lexicon. Two patterns are immediately apparent: positive adjectives describing women are often related to their bodies, while positive adjectives describing men are often related to their behavior. These patterns hold generally and the differences are significant (see §4).

\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|c|}
\hline
 & Female & & Male & \\
\hline
Positive & Negative & Positive & Negative \\
\hline
beautiful & battered & just & unsuitable \\
lovely & untreated & sound & unreliable \\
chaste & barren & righteous & lawless \\
gorgeous & shrewish & rational & inseparable \\
fertile & sheltered & peaceful & brutish \\
beauteous & heartbroken & prodigious & idle \\
sexy & unmarried & brave & unarmed \\
classy & undernourished & paramount & wounded \\
exquisite & underweight & reliable & bigoted \\
vivacious & uncomplaining & sinless & unjust \\
vibrant & nagging & honorable & brutal \\
\hline
\end{tabular}
\caption{Adjectives, with sentiment, used to describe men and women, as represented by our model. Colors indicate the most common sense of each adjective from Tsvetkov et al. (2014); black indicates out of lexicon. Two patterns are immediately apparent: positive adjectives describing women are often related to their bodies, while positive adjectives describing men are often related to their behavior. These patterns hold generally and the differences are significant (see §4).}
\end{table}

\textsuperscript{1}A noun’s natural gender is the implied gender of its referent (e.g., actress refers to woman). We distinguish natural
(or dependent) noun, abstracting away the noun form. To that end, we introduce a generative latent-variable model for representing gendered language, along with sentiment, from a parsed corpus. This model allows us to quantify differences between the language used to describe men and women.

The motivation behind our approach is straightforward: Consider the sets of adjectives (or verbs) that attach to gendered, animate nouns, such as man or woman. Do these sets differ in ways that depend on gender? For example, we might expect that the adjective Baltimorean attaches to man roughly the same number of times as it attaches to woman, controlling for the frequency of man and woman.² But this is not the case for all adjectives. The adjective pregnant, for example, almost always describes women, modulo the rare times that men are described as being pregnant with, say, emotion. Arguably, the gendered use of pregnant is benign—it is not due to cultural bias that women are more often described as pregnant, but rather because women bear children. However, differences in the use of other adjectives (or verbs) may be more pernicious. For example, female professors are less often described as brilliant than male professors (Storage et al., 2016), likely reflecting implicit or explicit stereotypes about men and women.

In this paper, we therefore aim to quantify the degree to which the language used to describe men and women is different and, moreover, different in a positive or negative way. Concretely, we focus on three sociolinguistic research questions about the influence of gender on adjective and verb choice:

Q1 What are the qualitative differences between the language used to describe men and women? For example, what, if any, are the patterns revealed by our model? Does the output from our model correlate with previous human judgments of gender stereotypes?

Q2 What are the quantitative differences between the language used to describe men and women? For example, are adjectives used to describe women more often related to their bodies than adjectives used to describe men? Can we quantify such patterns using existing semantic resources (Tsvetkov et al., 2014)?

Q3 Does the overall sentiment of the language used to describe men and women differ?

To answer these questions, we introduce a generative latent-variable model that jointly represents adjective (or verb) choice, with its sentiment, given the natural gender of a head (or dependent) noun. We use a form of posterior regularization to guide inference of the latent variables (Ganchev et al., 2010). We then use this model to study the syntactic n-gram corpus of (Goldberg and Orwant, 2013).

To answer Q1, we conduct an analysis that reveals differences between descriptions of male and female nouns that align with common gender stereotypes captured by previous human judgments. When using our model to answer Q2, we find that adjectives used to describe women are more often related to their bodies (significant under a permutation test with \( p < 0.03 \)) than adjectives used to describe men (see Fig. 1 for examples). This finding accords with previous research (Norberg, 2016). Finally, in answer to Q3, we find no significant difference in the overall sentiment of the language used to describe men and women.

### Table 1: Counts, in millions, of male and female nouns present in the corpus of Goldberg and Orwant (2013).

<table>
<thead>
<tr>
<th>Noun</th>
<th>Female</th>
<th>Male</th>
</tr>
</thead>
<tbody>
<tr>
<td>other</td>
<td>2.2</td>
<td>6.8</td>
</tr>
<tr>
<td>daughter</td>
<td>1.4</td>
<td>1.8</td>
</tr>
<tr>
<td>lady</td>
<td>2.4</td>
<td>2.1</td>
</tr>
<tr>
<td>wife</td>
<td>3.3</td>
<td>2.9</td>
</tr>
<tr>
<td>mother</td>
<td>4.2</td>
<td>4.2</td>
</tr>
<tr>
<td>girl</td>
<td>5.1</td>
<td>5.1</td>
</tr>
<tr>
<td>woman</td>
<td>11.5</td>
<td>39.9</td>
</tr>
<tr>
<td>Total</td>
<td>30.2</td>
<td>62.7</td>
</tr>
</tbody>
</table>

² Men are written about more often than women. Indeed, the corpus we use exhibits this trend, as shown in Tab. 1.

---

gender from grammatical gender because the latter does not necessarily convey anything meaningful about the referent.

Follow the methodology employed in previous sociolinguistic studies of gendered language, we use syntactic collocations to make definitive claims about gendered relationships between words. This approach stands in contrast to bag-of-words analyses, where information about gendered relationships must be indirectly inferred. By studying the
adjectives and verbs that attach to gendered, animate nouns, we are able to more precisely quantify the degree to which the language used to describe men and women is different. To date, much of the corpus-based sociolinguistics research on gendered language has focused on differences between the adjectives (or verbs) that modify each noun in a particular gendered pair of nouns, such as boy–girl or man–woman (e.g., Pearce (2008); Caldas-Coulthard and Moon (2010); Norberg (2016)). To assess the differences, researchers typically report top collocates\(^3\) for one word in the pair, exclusive of collocates for the other. This approach has the effect of restricting both the amount of available data and the claims that can be made regarding gendered nouns more broadly. In contrast, we focus on multiple noun pairs (including plural forms) simultaneously, modeling how the choice of adjective (or verb) depends on the natural gender of the head (or dependent) noun, abstracting away the noun form. As a result, we are able to make broader claims.

**The corpus of Goldberg and Orwant (2013).** To extract the adjectives and verbs that attach to gendered, animate nouns, we use the corpus of Goldberg and Orwant (2013), who ran a then-state-of-the-art dependency parser on 3.5 million digitalized books. We believe that the size of this corpus (11 billion words) makes our study the largest collocational study of its kind. Previous studies have used corpora of under one billion words, such as the British National Corpus (100 million words) (Pearce, 2008), the New Model Corpus (100 million words) (Norberg, 2016), and the Bank of English Corpus (450 million words) (Moon, Rosamund, 2014). By default, the corpus of Goldberg and Orwant (2013) is broken down by year, but we aggregate the data across years to obtain roughly 37 million noun–adjectives pairs, 41 million subj–verb pairs, and 14 million DOBJ–verb pairs. We additionally lemmatize each word. For example, the noun stewardesses is lemmatized to a set of lexical features consisting of the genderless lemma STEWARD and the morphological features +FEM and +PL. This parsing and lemmatization process is illustrated in Fig. 2.

**Quantitative evaluation.** Our study is also quantitative in nature: we test concrete hypotheses about differences between the language used to describe men and women. For example, we test whether

women are more often described using adjectives related to their bodies and emotions. This quantitative focus differentiates our approach from previous corpus-based sociolinguistics research on gendered language. Indeed, in the introduction to a special issue on corpus methods in the journal *Gender and Language*, Baker (2013) writes, “while the term corpus and its plural corpora are reasonably popular within Gender and Language (occurring in almost 40% of articles from issues 1-6), authors have mainly used the term as a synonym for ‘data set’ and have tended to carry out their analysis by hand and eye methods alone.” Moreover, in a related paper on extracting gendered language from word embeddings, Garg et al. (2018) lament that “due to the relative lack of systematic quantification of stereotypes in the literature [...] they] cannot directly validate [their] results.” For an overview of quantitative evaluation, we recommend Baker (2014).

**Speaker versus referent.** Many data-driven studies of gender and language focus on what speakers of different genders say rather than differences between descriptions of men and women. This is an easier task—the only annotation required is the gender of the speaker. For example, Ott (2016) used a topic model to study how word choice in tweets is influenced by the gender of the tweeter; Schofield and Mehr (2016) modeled gender in film dialog; and, in the realm of social media analysis, Bamman et al. (2014) discussed stylistic choices that enable classifiers to distinguish between tweets written by men versus women.

**Model versus data.** Recent NLP research has focused on gender biases in word embeddings (Bolukbasi et al., 2016; Zhao et al., 2017) and co-reference systems (Zhao et al., 2018; Rudinger et al., 2018). These papers are primarily concerned with mitigating biases present in the output of machine learning models deployed in the real world (O’Neil, 2016). For example, Bolukbasi et al. (2016) used pairs

---

\(^3\)Typically ranked by the log of the Dice coefficient.
of gendered words, such as she–he, to mitigate unwanted gender biases in word embeddings. Although it is possible to rank the adjectives (or verbs) most aligned with the embedding subspace defined by a pair of gendered words, there are no guarantees that the resulting adjectives (or verbs) were specifically used to describe men or women in the dataset from which the embeddings were learned. In contrast, we use syntactic collocations to explicitly represent gendered relationships between individual words. As a result, we are able make definitive claims about these relationships, thereby enabling us to answer sociolinguistic research questions. Indeed, it is this sociolinguistic focus that differentiates our approach from this line of work.

3 Modeling Gendered Language

As explained in §1, our aim is quantify the degree to which the language used to describe men and women is different and, moreover, different in a positive or negative way. To do this, we therefore introduce a generative latent-variable model that jointly represents adjective (or verb) choice, with its sentiment, given the natural gender of a head (or dependent) noun. This model, which is based on the sparse additive generative model (SAGE; Eisenstein et al., 2011), enables us to extract ranked lists of adjectives (or verbs) that are used, with particular sentiments, to describe male or female nouns.

We define \( \mathcal{G} \) to be the set of gendered, animate nouns in our corpus and \( n \in \mathcal{G} \) to be one such noun. We represent \( n \) via a multi-hot vector \( f_n \in \{0, 1\}^T \) of its lexical features—i.e., its genderless lemma, its gender (male or female), and its number (singular or plural). In other words, \( f_n \) always has exactly three non-zero entries; for example, the only non-zero entries of \( f_{\text{stewardess}} \) are those corresponding to \textit{STEWARD}, \textit{+FEM}, and \textit{+PL}. We define \( \mathcal{V} \) to be the set of adjectives (or verbs) in our corpus and \( \nu \in \mathcal{V} \) to be one such adjective (or verb). To simplify exposition, we refer to each adjective (or verb) that attaches to noun \( n \) as a neighbor of \( n \). Finally, we define \( \mathcal{S} = \{\text{POS}, \text{NEG}, \text{NEU}\} \) to be a set of three sentiments and \( s \in \mathcal{S} \) to be one such sentiment.

Drawing inspiration from SAGE, our model jointly represents nouns, neighbors, and (latent) sentiments as depicted in Fig. 3. Specifically,

\[
p(\nu, n, s) = p(\nu \mid s, n) p(s \mid n) p(n).
\]  

The first factor in eq. (1) is defined as

\[
p(\nu \mid s, n) \propto \exp \left( m_\nu + f_n^\top \eta(\nu, s) \right),
\]  

where \( m \in \mathbb{R}^{|\mathcal{V}|} \) is a background distribution and \( \eta(\nu, s) \in \mathbb{R}^T \) is a neighbor- and sentiment-specific deviation. The second factor in eq. (1) is defined as

\[
p(s \mid n) \propto \exp \left( \omega^n_s \right),
\]  

where \( \omega^n_s \in \mathbb{R} \), while the third factor is defined as

\[
p(n) \propto \exp \left( \xi_n \right),
\]  

where \( \xi_n \in \mathbb{R} \). We can then extract lists of neighbors that are used, with particular sentiments, to describe male and female nouns, ranked by scores that are a function of their deviations. For example, the score for neighbor \( \nu \) when used, with positive sentiment, to describe a male noun is defined as

\[
\tau_{\text{MASC-POS}}(\nu) \propto \exp \{g_{\text{MASC}}^\top \eta(\nu, \text{POS})\},
\]  

where \( g_{\text{MASC}} \in \{0, 1\}^T \) is a vector where only the entry that corresponds to \textit{+MASC} is non-zero.

Because our corpus does not contain explicit sentiment information, we marginalize out \( s \):

\[
p(\nu, n) = \sum_{s \in \mathcal{S}} p(\nu \mid s, n) p(s \mid n) p(n).
\]  

This yields the following objective function:

\[
\sum_{n \in \mathcal{G}} \sum_{\nu \in \mathcal{V}} \hat{p}(\nu, n) \log \left( p(\nu, n) \right),
\]  

where \( \hat{p}(\nu, n) \propto \#(\nu, n) \) is the empirical probability of neighbor \( \nu \) and noun \( n \) in our corpus.

To ensure that the latent variables in our model correspond to positive, negative, and neutral sentiments, we rely on posterior regularization (Ganchev et al., 2010). Given an additional distribution \( q(s \mid \nu) \) that provides external information
about the sentiment of neighbor \( \nu \), we regularize \( p(s \mid \nu) \), as defined by our model, to be close (in the sense of KL-divergence) to \( q(s \mid \nu) \). Specifically, we construct the following posterior regularizer:

\[
R_{post} = \text{KL}(q(s \mid \nu) \mid \mid p(s \mid \nu)) = -\sum_{s \in S} q(s \mid \nu) \log (p(s \mid \nu)) + H(q),
\]

where \( H(q) \) is constant and \( p(s \mid \nu) \) is defined as

\[
p(s \mid \nu) = \sum_{n \in G} p(s, n \mid \nu) = \sum_{n \in G} \frac{p(\nu \mid n, s) p(s \mid n) p(n)}{p(\nu)}.
\]

We use the combined sentiment lexicon of Hoyle et al. (2019) as \( q(s \mid \nu) \). This lexicon represents each word’s sentiment as a three-dimensional Dirichlet distribution, thereby accounting for the relative confidence in the strength of each sentiment and, in turn, accommodating polysemous and rare words. By using the lexicon as external information in our posterior regularizer, we can control the extent to which it influences the latent variables.

We add the regularizer in eq. (8) to the objective function in eq. (7), using a multiplier \( \beta \) to control the strength of the posterior regularization. We also impose an \( L_1 \)-regularizer \( \alpha \cdot ||\eta||_1 \) to induce sparsity. The complete objective function is then

\[
\sum_{n \in G} \sum_{\nu \in V} \beta(\nu \mid n, s) \log (p(\nu \mid n, s)) + \alpha \cdot ||\eta||_1 + \beta \cdot R_{post}.
\]

We optimize eq. (12) with respect to \( \eta(\cdot, \cdot), \omega, \) and \( \xi \) using the Adam optimizer (Kingma and Ba, 2015) with \( \alpha \) and \( \beta \) set as described in §4. To ensure that the parameters are interpretable (e.g., to avoid a negative \( \eta(\text{PREGNANT}, \text{NEG}) \) canceling out a positive \( \eta(\text{PREGNANT}, \text{POS}) \)), we also constrain \( \eta(\cdot, \cdot) \) to be non-negative, although without this constraint, our results are largely the same.

**Relationship to pointwise mutual information.** Our model also recovers pointwise mutual information (PMI), which has been used previously to identify gendered language (Rudinger et al., 2017).

**Proposition 1.** Consider the following restricted version of our model. Let \( f_g \in \{0,1\}^2 \) be a one-hot vector that represents only the gender of a noun. We write \( g \) instead of \( n \), equivalence-classing all nouns as either MASC or FEM. Let \( \eta^*(\cdot) : V \rightarrow \mathbb{R}^2 \) be the maximum-likelihood estimate for the special case of our model without (latent) sentiments:

\[
p(\nu \mid g) \propto \exp(m_\nu + f_g^T \eta^*(\nu)).
\]

Then, we have

\[
\tau_g(\nu) \propto \exp(\text{PMI}(\nu, g)).
\]

**Proof.** See App. B.

Proposition 1 says that if we use a limited set of lexical features (i.e., only gender) and estimate our model without any regularization or latent sentiments, then ranking the neighbors by \( \tau_g(\nu) \) (i.e., by their deviations from the background distribution) is equivalent to ranking them by their PMI. This proposition therefore provides insight into how our model builds on PMI. Specifically, in contrast to PMI, 1) our model can consider lexical features other than gender, 2) our model is regularized to avoid the pitfalls of maximum-likelihood estimation, and 3) our model cleanly incorporates latent sentiments, relying on posterior regularization to ensure that the \( p(s \mid \nu) \) is close to the sentiment lexicon of Hoyle et al. (2019).

## 4 Experiments, Results, and Discussion

We use our model to study the corpus of Goldberg and Orwant (2013) by running it separately on the noun–adjectives pairs, the NSUBJ–verb pairs, and the DOBJ–verb pairs. We provide a full list of the lemmatized, gendered, animate nouns in App. A. We use \( \alpha \in \{0, 10^{-5}, 10^{-4}, 0.001, 0.01\} \) and \( \beta \in \{10^{-5}, 10^{-4}, 0.001, 0.01, 0.1, 1, 10, 100\} \); when we report results below, we use parameter values averaged over these hyperparameter settings.

### 4.1 Q1: Qualitative Differences

Our first research question concerns the qualitative differences between the language used to describe men and women. To answer this question, we use our model to extract ranked lists of neighbors that are used, with particular sentiments, to describe male and female nouns. As explained in §3, we rank the neighbors by their deviations from the background distribution (see, for example, eq. (5)).

**Qualitative evaluation.** In Tab. 2, we provide, for each sentiment, the 25 largest-deviation adjectives used to describe male and female nouns. The
results are striking: it is immediately apparent that positive adjectives describing women are often related to their appearance (e.g., beautiful, fair, and pretty). Sociolinguistic studies of other corpora, such as British newspapers (Caldas-Coulthard and Moon, 2010), have also revealed this pattern. Adjectives relating to fertility, such as fertile and barren, are also more prevalent for women. We provide similar tables for verbs in App. D. Negative verbs describing men are often related to violence (e.g., murder, fight, kill, and threaten). Meanwhile, women are almost always the object of rape, which aligns with our knowledge of the world and supports the collocation of rape and girl found by Baker (2014). Broadly speaking, positive verbs describing men tend to connote virtuosity (e.g., gallant and inspire), while those describing women appear more trivial (e.g., sprightly, giggle, and kiss).

**Correlation with human judgments.** To determine whether the output from our model accords with previous human judgements of gender stereotypes, we use the corpus of Williams and Bennett (1975), which consists of 63 adjectives annotated with (binary) gender stereotypes. We measure Spearman’s $\rho$ between these annotations and the probabilities output by our model. We find a relatively strong positive correlation of $\rho = 0.59$ ($p < 10^{-6}$), which indicates that the output from our model aligns with common gender stereotypes captured by previous human judgements. We also measure the correlation between continuous annotations of 300 adjectives from two follow-up studies (Williams and Best, 1990, 1977)\(^3\) and the probabilities output by our model. Here, the correlation is $\rho = 0.33$ ($p < 10^{-8}$), and the binarized annotations agree with the output from our model for 64% of terms. We note that some of the disagreement is due to reporting bias (Gordon and Van Durme, 2013) in our corpus. For example, only men are described in our corpus as effeminate, although humans judge it to be a highly feminine adjective.

### 4.2 Q2: Quantitative differences

Our second research question concerns the quantitative differences between the language used to describe men and women. To answer this question, we use two existing semantic resources—one for adjectives (Tsvetkov et al., 2014) and one for verbs (Miller et al., 1993)—to quantify the patterns revealed by our model. Again, we use our model to extract ranked lists of neighbors that are used, with particular sentiments, to describe male and female nouns. We consider only the 200 largest-deviation

\(^3\)The studies consider the same set of words 20 years apart; we average their annotations, obtained from Garg et al. (2018).
neighbours for each sentiment and gender. This restriction allows us to perform an unpaired permutation test (Good, 2004) to determine whether there are significant differences between the language used to describe men and women.

**Adjective evaluation.** Women are supposedly more often described using adjectives related to their bodies and emotions. For example, de Beauvoir (1953) writes that “from girlhood, women are socialized to live and experience their bodies as objects for another’s gaze...” Although studies of reasonably large corpora have found evidence to support this supposition (Norberg, 2016), none have done so at scale with statistical significance testing. We use the semantic resource of Tsvetkov et al. (2014), which categorizes adjectives into thirteen senses: BEHAVIOR, BODY, FEELING, MIND, etc. Specifically, each adjective has a distribution over senses, capturing how often the adjective corresponds to each sense. We analyze the largest-deviation adjectives for each sentiment and gender by computing the frequency with which these adjectives correspond to each sense. We depict these frequencies in Fig. 4. Specifically, we provide frequencies for the senses where, after Bonferroni correction, the differences between men and women are significant. We find that adjectives used to describe women are indeed more often related to their bodies and emotions than adjectives used to describe men.

**Verb evaluation.** To evaluate verbs senses, we take the same approach as for adjectives. We use the semantic resource of Miller et al. (1993), which categorizes verbs into fifteen senses. Each verb has a distribution over senses, capturing how often the verb corresponds to each sense. We consider two cases: the NSUBJ–verb pairs and the DOBJ–verb pairs. Overall, there are fewer significant differences for verbs than there are for adjectives. There are no statistically significant differences for the DOBJ–verb pairs. We depict the results for the NSUBJ–verb pairs in Fig. 5. We find that verbs used to describe women are more often related to their bodies than verbs used to describe men.

### 4.3 Q3: Differences in sentiment

Our final research question concerns the overall sentiment of the language used to describe men and women. To answer this question, we use a simplified version of our model, without the latent sentiment variables or the posterior regularizer. We are then able to use the combined sentiment lexicon of Hoyle et al. (2019) to analyze the largest-deviation

<table>
<thead>
<tr>
<th></th>
<th>ADJ</th>
<th>NSUBJ</th>
<th>DOBJ</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MSC</td>
<td>FEM</td>
<td>MSC</td>
</tr>
<tr>
<td>POS</td>
<td>0.34</td>
<td>0.38</td>
<td>0.37</td>
</tr>
<tr>
<td>NEG</td>
<td>0.30</td>
<td>0.31</td>
<td>0.33</td>
</tr>
<tr>
<td>NEU</td>
<td><strong>0.36</strong></td>
<td><strong>0.31</strong></td>
<td>0.30</td>
</tr>
</tbody>
</table>

Table 3: The frequency with which the 200 largest-deviation neighbors for each gender correspond to each sentiment, obtained using a simplified version of our model and the lexicon of Hoyle et al. (2019). Significant differences ($p < 0.05/3$ under an unpaired permutation test with Bonferroni correction) are in bold.
neighbors for each gender by computing the frequency with which each neighbor corresponds to each sentiment. We report these frequencies in Tab. 3. We find that there is only one significant difference: adjectives used to describe men are more often neutral than those used to describe women.

5 Conclusion and Limitations

We presented an experimental framework for quantitatively studying the ways in which the language used to describe men and women is different and, moreover, different in a positive or negative way. We introduced a generative latent-variable model that jointly represents adjective (or verb) choice, with its sentiment, given the natural gender of a head (or dependent) noun. Via our experiments, we found evidence in support of common gender stereotypes. For example, positive adjectives used to describe women are more often related to their bodies than adjectives used to describe men. Our study has a few limitations that we wish to highlight. First, we ignore demographics (e.g., age, gender, location) of the speaker, even though such demographics are likely influence word choice. Second, we ignore genre (e.g., news, romance) of the text, even though genre is also likely to influence the language used to describe men and women. In addition, depictions of men and women have certainly changed over the period covered by our corpus; indeed, Underwood et al. (2018) found evidence of such a change for fictional characters. In future work, we intend to conduct a diachronic analysis in English using the same corpus, in addition to a cross-linguistic study of gendered language.

Acknowledgments

We would like to thank the three anonymous ACL 2019 reviewers for their comments on the submitted version, as well as the anonymous reviewers of a previous submission. We would also like to thank Adina Williams and Eleanor Chodroff for their comments on versions of the manuscript. The last author would like to acknowledge a Facebook fellowship.

References


Simone de Beauvoir. 1953. The second sex.


Margaret Ott. 2016. Tweet like a girl: Corpus analysis of gendered language in social media.


John E. Williams and Susan M. Bennett. 1975. The definition of sex stereotypes via the adjective check list. Sex Roles, 1(4):327–337.


A List of Gendered, Animate Nouns

Tab. 4 contains the full list of gendered, animate nouns that we use. We consider each row in this table to be the inflected forms of a single lemma.

<table>
<thead>
<tr>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>Singular</td>
<td>Plural</td>
</tr>
<tr>
<td>man</td>
<td>men</td>
</tr>
<tr>
<td>boy</td>
<td>boys</td>
</tr>
<tr>
<td>father</td>
<td>fathers</td>
</tr>
<tr>
<td>son</td>
<td>sons</td>
</tr>
<tr>
<td>brother</td>
<td>brothers</td>
</tr>
<tr>
<td>husband</td>
<td>husbands</td>
</tr>
<tr>
<td>uncle</td>
<td>uncles</td>
</tr>
<tr>
<td>nephew</td>
<td>nephews</td>
</tr>
<tr>
<td>emperor</td>
<td>emperors</td>
</tr>
<tr>
<td>king</td>
<td>kings</td>
</tr>
<tr>
<td>prince</td>
<td>princes</td>
</tr>
<tr>
<td>duke</td>
<td>dukes</td>
</tr>
<tr>
<td>lord</td>
<td>lords</td>
</tr>
<tr>
<td>knight</td>
<td>knights</td>
</tr>
<tr>
<td>waiter</td>
<td>waiters</td>
</tr>
<tr>
<td>actor</td>
<td>actors</td>
</tr>
<tr>
<td>god</td>
<td>gods</td>
</tr>
<tr>
<td>policemman</td>
<td>policemen</td>
</tr>
<tr>
<td>postman</td>
<td>postmen</td>
</tr>
<tr>
<td>hero</td>
<td>heros</td>
</tr>
<tr>
<td>wizard</td>
<td>wizards</td>
</tr>
<tr>
<td>steward</td>
<td>stewards</td>
</tr>
<tr>
<td>he</td>
<td>–</td>
</tr>
</tbody>
</table>

Table 4: Gendered, animate nouns.

B Relationship to PMI

Proposition 1. Consider the following restricted version of our model. Let \( f_g \in \{0, 1\}^2 \) be a one-hot vector that represents only the gender of a noun. We write \( g \) instead of \( n \), equivalence-classing all nouns as either MASC or FEM. Let \( \eta^*(\cdot) : \mathcal{V} \rightarrow \mathbb{R}^2 \) be the maximum-likelihood estimate for the special case of our model without (latent) sentiments:

\[
p(\nu | g) \propto \exp(m_\nu + f_g^\top \eta^*(\nu)).
\] (15)

Then, we have

\[
\tau_g(\nu) \propto \exp(\text{PMI}(\nu, g)).
\] (16)

Proof. First, we note our model has enough parameters to fit the empirical distribution exactly:

\[
\hat{p}(\nu | g) = p(\nu | g)
\] (17)

\[
\propto \exp\{m_\nu + f_g^\top \eta^*(\nu)\}.
\] (18)

Then, we proceed with an algebraic manipulation of the definition of pointwise mutual information:

\[
\text{PMI}(\nu, g) = \log \frac{\hat{p}(\nu, n)}{\hat{p}(\nu) \hat{p}(n)} = \log \frac{\hat{p}(\nu | n)}{\hat{p}(\nu)} = \log \frac{p(\nu | n)}{\hat{p}(\nu)} = \log \frac{p(\nu | n)}{\exp\{m_\nu\}}
\] (19)

\[
= \log \frac{1}{Z} \exp\{m_\nu + f_g^\top \eta^*(\nu)\} = \log \frac{1}{Z} \exp\{f_g^\top \eta^*(\nu)\} = f_g^\top \eta^*(\nu) - \log Z.
\] (20)

Now we have

\[
\tau_g(\nu) \propto \exp\{f_g^\top \eta^*(\nu)\} \propto \exp\{f_g^\top \eta^*(\nu) - \log Z\} = \exp(\text{PMI}(\nu, g)),
\] (21)

which is what we wanted to show.

C Senses

In Tab. 5, we list the senses for adjectives (Tsvetkov et al., 2014) and for verbs (Miller et al., 1993).

<table>
<thead>
<tr>
<th>Adjectives</th>
<th>Verbs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Behavior</td>
<td>Body</td>
</tr>
<tr>
<td>Body</td>
<td>Change</td>
</tr>
<tr>
<td>Feeling</td>
<td>Cognition</td>
</tr>
<tr>
<td>Mind</td>
<td>Communication</td>
</tr>
<tr>
<td>Miscellaneous</td>
<td>Competition</td>
</tr>
<tr>
<td>Motion</td>
<td>Consumption</td>
</tr>
<tr>
<td>Perception</td>
<td>Contact</td>
</tr>
<tr>
<td>Quantity</td>
<td>Creation</td>
</tr>
<tr>
<td>Social</td>
<td>Emotion</td>
</tr>
<tr>
<td>Spatial</td>
<td>Motion</td>
</tr>
<tr>
<td>Substance</td>
<td>Perception</td>
</tr>
<tr>
<td>Temporal</td>
<td>Possession</td>
</tr>
<tr>
<td>Weather</td>
<td>Social</td>
</tr>
<tr>
<td>Stative</td>
<td>Weather</td>
</tr>
</tbody>
</table>

Table 5: Senses for adjectives and verbs.

D Additional Results

In Tab. 6 and Tab. 7, we provide the largest-deviation verbs used to describe male and female nouns for NSUBJ–verb pairs and DOBJ–verb pairs.
<table>
<thead>
<tr>
<th>Verb</th>
<th>Value</th>
<th>Verb</th>
<th>Value</th>
<th>Verb</th>
<th>Value</th>
<th>Verb</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>succeed</td>
<td>1.6</td>
<td>fight</td>
<td>1.2</td>
<td>extend</td>
<td>0.7</td>
<td>celebrate</td>
<td>2.4</td>
</tr>
<tr>
<td>protect</td>
<td>1.4</td>
<td>fail</td>
<td>1.0</td>
<td>found</td>
<td>0.8</td>
<td>fascinate</td>
<td>0.8</td>
</tr>
<tr>
<td>favor</td>
<td>1.3</td>
<td>fear</td>
<td>1.0</td>
<td>strike</td>
<td>1.3</td>
<td>facilitate</td>
<td>0.7</td>
</tr>
<tr>
<td>flourish</td>
<td>1.3</td>
<td>murder</td>
<td>1.5</td>
<td>own</td>
<td>1.1</td>
<td>marry</td>
<td>1.8</td>
</tr>
<tr>
<td>prosper</td>
<td>1.7</td>
<td>shock</td>
<td>1.6</td>
<td>collect</td>
<td>1.1</td>
<td>smile</td>
<td>1.8</td>
</tr>
<tr>
<td>support</td>
<td>1.5</td>
<td>blind</td>
<td>1.6</td>
<td>set</td>
<td>0.8</td>
<td>fan</td>
<td>0.8</td>
</tr>
<tr>
<td>promise</td>
<td>1.5</td>
<td>forbid</td>
<td>1.5</td>
<td>wag</td>
<td>1.0</td>
<td>kiss</td>
<td>1.8</td>
</tr>
<tr>
<td>welcome</td>
<td>1.5</td>
<td>kill</td>
<td>1.3</td>
<td>present</td>
<td>0.9</td>
<td>champion</td>
<td>2.2</td>
</tr>
<tr>
<td>favour</td>
<td>1.2</td>
<td>protest</td>
<td>1.3</td>
<td>pretend</td>
<td>1.1</td>
<td>adore</td>
<td>2.0</td>
</tr>
<tr>
<td>clear</td>
<td>1.9</td>
<td>cheat</td>
<td>1.3</td>
<td>prostate</td>
<td>1.1</td>
<td>dance</td>
<td>1.7</td>
</tr>
<tr>
<td>reward</td>
<td>1.8</td>
<td>fake</td>
<td>0.8</td>
<td>want</td>
<td>0.9</td>
<td>laugh</td>
<td>1.6</td>
</tr>
<tr>
<td>appeal</td>
<td>1.6</td>
<td>deprive</td>
<td>1.5</td>
<td>create</td>
<td>0.9</td>
<td>have</td>
<td>1.4</td>
</tr>
<tr>
<td>encourage</td>
<td>1.5</td>
<td>threaten</td>
<td>1.3</td>
<td>pay</td>
<td>1.1</td>
<td>play</td>
<td>1.0</td>
</tr>
<tr>
<td>allow</td>
<td>1.5</td>
<td>frustrate</td>
<td>0.9</td>
<td>prompt</td>
<td>1.0</td>
<td>give</td>
<td>0.8</td>
</tr>
<tr>
<td>respect</td>
<td>1.5</td>
<td>fright</td>
<td>0.9</td>
<td>brazen</td>
<td>1.0</td>
<td>like</td>
<td>1.8</td>
</tr>
<tr>
<td>comfort</td>
<td>1.4</td>
<td>temper</td>
<td>1.4</td>
<td>tarry</td>
<td>0.7</td>
<td>giggle</td>
<td>1.4</td>
</tr>
<tr>
<td>treat</td>
<td>1.3</td>
<td>horror</td>
<td>1.4</td>
<td>front</td>
<td>0.5</td>
<td>extol</td>
<td>0.6</td>
</tr>
<tr>
<td>brave</td>
<td>1.7</td>
<td>neglect</td>
<td>1.4</td>
<td>flush</td>
<td>0.3</td>
<td>compassionate</td>
<td>1.9</td>
</tr>
<tr>
<td>rescue</td>
<td>1.5</td>
<td>argue</td>
<td>1.3</td>
<td>reach</td>
<td>0.9</td>
<td>live</td>
<td>1.4</td>
</tr>
<tr>
<td>win</td>
<td>1.5</td>
<td>denounce</td>
<td>1.3</td>
<td>escape</td>
<td>0.8</td>
<td>free</td>
<td>0.9</td>
</tr>
<tr>
<td>warm</td>
<td>1.5</td>
<td>concern</td>
<td>1.2</td>
<td>gi</td>
<td>0.7</td>
<td>felicitate</td>
<td>0.6</td>
</tr>
<tr>
<td>praise</td>
<td>1.4</td>
<td>expel</td>
<td>1.7</td>
<td>rush</td>
<td>0.6</td>
<td>mature</td>
<td>2.2</td>
</tr>
<tr>
<td>fit</td>
<td>1.4</td>
<td>dispute</td>
<td>1.5</td>
<td>duplicate</td>
<td>0.5</td>
<td>exalt</td>
<td>1.7</td>
</tr>
<tr>
<td>wish</td>
<td>1.4</td>
<td>obscure</td>
<td>1.4</td>
<td>incarnate</td>
<td>0.5</td>
<td>surpass</td>
<td>1.7</td>
</tr>
<tr>
<td>grant</td>
<td>1.3</td>
<td>damn</td>
<td>1.4</td>
<td>freeze</td>
<td>0.5</td>
<td>meet</td>
<td>1.1</td>
</tr>
</tbody>
</table>

Table 6: The largest-deviation verbs used to describe male and female nouns for NSUBJ–verb pairs.