Linguistic Bias in Collaboratively Produced Biographies:
Crowdsourcing Social Stereotypes?

Jahna Otterbacher
Graduate Program in Social Information Systems
Open University of Cyprus
jahna.otterbacher@ouc.ac.cy

Abstract

Language is the primary medium through which stereotypes are conveyed. Even when we avoid using derogatory language, there are many subtle ways in which stereotypes are created and reinforced, and they often go unnoticed. Linguistic bias, the systematic asymmetry in language patterns as a function of the social group of the persons described, may play a key role. We ground our study in the social psychology literature on linguistic biases, and consider two ways in which biases might manifest: through the use of more abstract versus concrete language, and subjective words. We analyze biographies of African American and Caucasian actors at the Internet Movie Database (IMDb), hypothesizing that language patterns vary as a function of race and gender. We find that both attributes are correlated to the use of abstract, subjective language. Theory predicts that we describe people and scenes that are expected, as well as positive aspects of our in-group members, with more abstract language. Indeed, white actors are described with more abstract, subjective language at IMDb, as compared to other social groups. Abstract language is powerful because it implies stability over time; studies have shown that people have better impressions of others described in abstract terms. Therefore, the widespread prevalence of linguistic biases in social media stands to reinforce social stereotypes. Further work should consider the technical and social characteristics of the collaborative writing process that lead to an increase or decrease in linguistic biases.

Introduction

The rise of social media has provided us a variety of means to offer cognitive surplus in the creation and sharing of knowledge that can benefit everyone (Shirkey 2011). From collaborative efforts such as Wikipedia, to systems that aggregate individual contributions (e.g., consumer-contributed reviews), the common goal is to document collective wisdom about a growing number of subjects. With information-centric social media among the most popular on the Web (e.g., Amazon.com ranked sixth and Wikipedia.org ranked seventh among Websites worldwide by Alexa at the time of writing\(^1\), it is clear that what gets published in such systems is of growing importance in society. The information shared represents “what is known” or “what we believe to be true” at the collective level, at a given point in time.

The collaboratively produced biography is a genre of social media text receiving increasing attention. Such biographies can be found at general-purpose, open-edit knowledge production and sharing sites (e.g., Wikipedia\(^2\)), as well as at special-interest community sites (e.g., the Internet Movie Database (IMDb)\(^3\) or Music Wiki\(^4\)). In these systems, users create and edit biographies. The accuracy, completeness and clarity of the information provided often reach a high level, given the participation of sufficient numbers of contributors (Kittur and Kraut 2008). In fact, the introduction of incorrect information is often corrected within a matter of hours\(^5\). However, it is not only information that is inaccurate or misleading that should concern us, but also the subtle communication patterns that create or reinforce social stereotypes.

Quality and Bias in Collaborative Biographies

It is not surprising that the quality of collaboratively produced biographies of famous people has been the focus of previous research. For those still living, one’s digital biographies convey his or her reputation, and the consequences this reputation can have on one’s life experiences are obvious. For the deceased, crowdsourced digital biographies serve as a collective memory (Pentzold 2009) of the person, her lifetime accomplishments and how she was as a character. Flekova and colleagues (Flekova, Ferschke, and Gurevych 2014) reported that over one-fifth of Wikipedia articles describes people, most of whom are still living. They call for the development of automated methods to ensure that biographies are of a high quality, given their vulnerability to corruption and vandalism. Developing machine learning techniques guided by human judgments on quality, they scored articles on four dimensions: completeness, writing quality, trustworthiness and, most relevant to our work,
objectivity. Interestingly, they found that textual features (rather than Wikipedia-based features such as links or article age), including the use of words carrying sentiment, were the best predictors of biography subjectivity/objectivity. This demonstrates that the manner in which we describe others conveys information not only about the subject described, but also about ourselves as information sources.

In a similar vein, others have described biases in collaboratively produced knowledge resources. Wikipedia again receives the lion’s share of attention, with previous work examining whether the user base is really diverse (i.e., participation biases) and how information is documented (i.e., content biases). As mentioned, with a sufficiently numerous and diverse group of participants, collaboratively produced texts reach a high level of quality. However, in online communities, a particular user demographic, men, often dominates. At Wikipedia, men hold the majority both in terms of active users and edits, with many suggesting that Wikipedia essentially conveys a man’s view of the world (Lam et al. 2011; Antin et al. 2011; Forte et al. 2012). While it is difficult to assess how participation bias impacts the content created, it surely results in undesirable consequences, such as imbalances in the topics discussed and documented, and the development of community norms that women and minority social groups find intimidating (Herring 2003; Hemphill and Otterbacher 2012).

With respect to content biases at Wikipedia, across language versions, there is significant variation with respect to which topics are covered, thus reflecting cultural differences (Hecht and Gergle 2010). Callahan and Herring found significant differences in famous persons’ biographies, both as a function of the Wikipedia community (Polish vs. English) and the nationality of the person being described (Polish vs. American scientists) (Callahan and Herring 2011). Via content analysis, they discovered that more personal details were described for Americans, and that descriptions were more positive toward Americans as compared to the Polish scientists. In general, English language biographies were more positive in tone than those documented in Polish. Like any technology, Wikipedia is not values-free; what is covered there reflects the interests, culture, and social standing of the people who use it (Pfeil, Zaphiris, and Ang 2006; Royal and Kapila 2009).

Like the researchers cited here, we share a concern for the quality of collaboratively produced knowledge resources, and in particular, for biographies of persons both living and deceased. However, our work departs from previous research in important ways. We offer the first study of linguistic biases in social media descriptions of people. As will be explained, linguistic biases are not attributed to cultural differences, nor to any conscious attempt to alter the content created. They are believed to have cognitive origins, although as we will see, they have very social consequences.

Another point of departure is that we study biographies at the Internet Movie Database. IMDb is arguably a less formal knowledge source as compared to Wikipedia. It is not an encyclopedia, but rather, a database, aiming to be the “world’s most popular and authoritative source for movie, TV and celebrity content.” 6 It has no Neutral Point of View policy, does not feature a talk page, where users discuss their views and opinions on a page’s content and style, nor can one see which user edited what. Nonetheless, true to its mission, IMDb is currently ranked 49th in the world in terms of traffic, and 25th in the U.S., typically appearing in the first hits from a search engine, given a famous actor’s name. Motivated by theories of linguistic bias, we consider the language in biographies of famous African American and Caucasian actors and actresses, comparing the extent to which patterns vary based on the subject’s race and gender.

The Language of Biographies: Subtle Stereotypes?

Psychologists and communication scientists have become increasingly convinced that the manner in which we use language plays a key role in the transmission and maintenance of social stereotypes (Maass 1999; von Hippel, Sekaquaptewa, and Vargas 1997). Even when we avoid using derogatory language (e.g., racial slurs, sexist terms), our linguistic biases may still give away the stereotypes that influence us. We use Beukeboom’s definition of linguistic bias (Beukeboom 2013):

A systematic asymmetry in the way that one uses language, as a function of the social group of the person(s) being described.

We consider two patterns that may reveal underlying social expectations: the use of abstract versus concrete language, and the use of subjective words (e.g., nice, bad, beautiful, ugly). Consider the following three statements:

1. Morgan Freeman played in The Shawshank Redemption.
2. Morgan Freeman was amazing in Shawshank Redemption.
3. Morgan Freeman is an amazing actor.

The first statement is the most concrete and objective of the three: it is restricted to a particular context, and is devoid of subjective words. In contrast, the third sentence is the most abstract of the three, as it makes a general statement about Freeman (i.e., is not restricted to the context of a single film) and contains a subjective adjective (“amazing”). The question at hand is the extent to which we observe differences in the use of abstract and subjective language, as a function of the social groups of the actors.

Research Questions

Due to the growing number of media that enable us to collaboratively produce biographies, an important question is whether subtle linguistic biases make their way into these influential and popular texts. Linguistic biases have been explored extensively by social psychologists, although in small-scale, experimental contexts. However, we are unaware of previous studies that have attempted to understand

6 http://www.nytimes.com/roomfordebate/2011/02/02/where-are-the-women-in-wikipedia
7 http://www.imdb.com/pressroom/?ref_=ft_pr
Linguistic Biases

We describe two linguistic biases discussed by social psychologists and communication scientists: the Linguistic Expectancy Bias (LEB) and the Linguistic Intergroup Bias (LIB). These biases manifest through two characteristics of the language used to describe someone: the specificity of the description, and the use of words that reveal sentiment toward the target individual. Therefore, we begin with an overview of Semin and Fiedler’s Linguistic Category Model (LCM) (Semin and Fiedler 1988).

Linguistic Category Model

Both the LEB and the LIB build upon the Linguistic Category Model. LCM proposes a shift in the methodological approach to the analysis of language, from the individual to the social, emphasizing that “to understand social behavior one has to develop a handle on language as a tool that carries communication and makes social interaction possible,” ((Coenen, Hedebo, and Semin 2006), p. 4).

LCM specifies four categories of predicates with respect to the level of abstraction in a description of a person, as depicted in Figure 1. As illustrated, the most concrete description is that involving a descriptive action verb; it describes an observed event with no interpretation. In contrast, the most abstract is that involving an adjective; here, the description generalizes across any scenario or event. In between the two extremes, we have the use of a state verb, which describes an ongoing state of affairs, as well as the use of an interpretive action verb, in which what is being described is attributed only to a specific event or action.

As will be described, systematic differences in the level of abstractness of the language used to describe people, is used to detect LEB and LIB. The underpinnings of these biases are cognitive in nature, as familiar and/or stereotypical scenes are easier to process (Winkielman et al. 2006). However, the consequences are of a social nature and are quite serious, especially if they prove to be as pervasive in social media as they are in interpersonal interactions.

Abstract language is powerful because it implies stability over time, as well as generalizability across situations. It has been shown that recipients of messages are impacted by biases; they interpret abstract descriptions as enduring qualities of the target person, and concrete descriptions as being transient (Wigboldus, Spears, and Semin 2000). Thus, it is believed that linguistic biases contribute to the maintenance and transmission of stereotypes, as information encoded in an abstract manner is more resistant to disconfirmation.

Linguistic Expectancy Bias

LEB describes the tendency to describe other people and situations that are expectancy consistent (e.g., stereotype-congruent individuals) in a more abstract, interpretive way. Abstract descriptions of a target individual provide more information about their perceived traits and characteristics, and less about a particular situation they are in or action they have taken. Studies have shown that when a target individual violates our expectations, we are likely to focus on more tangible, concrete details in our descriptions of him or her (Maass et al. 1989; Wigboldus, Spears, and Semin 2005).

In contrast, stereotype-congruent people and behaviors are likely to be described more abstractly, using language that makes reference to their general disposition and traits. Although the LEB is pervasive in human communication, it has only been studied in the laboratory, with very few exceptions (e.g., (Hunt 2011)).

Linguistic Intergroup Bias

The key thesis of LIB is that we use language in a manner that renders the disconfirmation of preexisting ideas we hold about social groups very difficult (Maass et al. 1989). For members of our in-group, we tend to describe positive actions and attributes using more abstract language, and their undesirable behaviors and attributes more concretely. Conversely, when an out-group individual does or is something desirable, we tend to describe him or her with more concrete language, whereas her undesirable attributes are encoded more abstractly. LIB builds on LEB, since we expect our in-group members to have desirable attributes and/or to exhibit desirable behaviors, whereas positive characteristics of out-group members may be unexpected.

Linguistic biases are not only driven by cognitive and motivational underpinnings, they are also influenced by social and communicative context (Maass et al. 1989). Specifically, the LIB is more likely to occur when messages are designed to serve a clear communicative purpose. In other words, LIB

<table>
<thead>
<tr>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adjectives</td>
<td>Describes a characteristic or feature of a person.</td>
</tr>
<tr>
<td>State verb</td>
<td>Describes an enduring cognitive or emotional state with no clear beginning or end.</td>
</tr>
<tr>
<td>Interpretive action verb</td>
<td>Refers to various actions with clear beginning and end.</td>
</tr>
<tr>
<td>Descriptive action verb</td>
<td>Refers to a single, specific action with a clear beginning and end.</td>
</tr>
</tbody>
</table>

Figure 1: The Linguistic Category Model.

RQ1: Which types of linguistic biases could manifest in a textual biography of a famous person?

RQ2: Is there evidence of linguistic bias based on the gender of the person being described?

RQ3: Is there evidence of linguistic bias based on the race of the person described?
may serve as a device that signals to others both our status with respect to an in-/out-group, as well as our expectations for their behaviors. When communicative purpose is removed from the situation (i.e., when people cannot construct an image of their audience), LIB is less likely to occur.

### Detecting Linguistic Bias in Social Media

Figure 2 summarizes the linguistic properties of textual biographies predicted by theory, as a function of the social relationship between biographers (i.e., IMDb participants) and the actors being described. As mentioned, IMDb is a less formal collaborative system as compared to Wikipedia, as it lacks many of the social features that build in reputation and accountability. These properties make it an interesting case for our study. Because linguistic biases are mitigated by the communicative context, we might expect collaborative biographies created in a more anonymous communication environment, such as IMDb, to suffer less from linguistic bias, where the social identity of the biography’s subject is the primary trigger for LIB and LEB.

Previous studies of linguistic bias have involved manually annotating textual descriptions of people by LCM categories (Semin and Fiedler 1988). While the LCM is clearly a complicated model to fully automate, we currently take the first steps toward this process. This will allow us to study a sufficient number of IMDb biographies in order to compare them across race and gender.

The LCM Manual notes that the textual segments we should annotate and how we should apply LCM depends on the research questions to be addressed (p. 8). Two clear observations emerge from our review of the literature. The first is that adjectives play a key role in conveying abstract information about people, and they can be distinguished from verbs in a straightforward manner (p. 6). Furthermore, linguists have long considered the possibility that part-of-speech (POS) is intricately linked to underlying cognitive functions (Brown 1957). Thus, the use of more adjectives over verbs likely affects how communicators view a situation or person.

The second observation is that subjective words of all POS play a key role in more abstract descriptions. Subjective language injects the author’s sentiment and/or inferences about the target person, into the description. Since our work is inspired by the claim that language biases play a role in the creation and maintenance of stereotypes, it is worth noting that many seminal works on stereotypes (e.g., (Devine and Elliot 1995)) ask participants to describe the core attributes of a social group of interest using subjective adjectives. Based on these insights, we consider the following textual features of biographies:

1. The use of adjectives versus verbs.
2. The use of subjective words of any POS.
3. The use of subjective adjectives.

We measure the above properties in various ways, and using appropriate statistical models, compare their use in IMDb biographies across actor race and gender. Significant differences across social groups will be a strong indication of linguistic bias.

### Data

Given that IMDb is based in the United States, we defined lists of top Hollywood actors. To this end, we used Wikipedia to identify names of prominent African American and Caucasian American actors and actresses. We then obtained the most recent dump of the IMDb biographies dataset. Table 1 summarizes the number of biographies from our list of prominent actors and actresses, by social group, that were available at IMDb.

### Preprocessing

From the 295 biographies we created three datasets: 1) the full text biographies, 2) the first five sentences of the biographies and 3) the opening sentence of each biography. The five-sentence dataset was used to emulate the “teaser” a reader sees at an actor’s page at IMDb, before clicking to see the entire biography. The first-sentence dataset was created with the intuition that it sets the general topic and tone of the text. Table 2 provides examples of the first sentences in four biographies, along with the proportion of words that are

<table>
<thead>
<tr>
<th>Traits</th>
<th>Expectancy-congruent (LEB)</th>
<th>Expectancy-incongruent (LEB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Familiar/Desirable</td>
<td>More abstract</td>
<td>More concrete</td>
</tr>
<tr>
<td>Traits</td>
<td>More adjectives</td>
<td>Fewer adjectives</td>
</tr>
<tr>
<td></td>
<td>More subjective words</td>
<td>Fewer subjective words</td>
</tr>
<tr>
<td>Unfamiliar/Undesirable</td>
<td>More concrete</td>
<td>More abstract</td>
</tr>
<tr>
<td>Traits</td>
<td>Fewer adjectives</td>
<td>More adjectives</td>
</tr>
<tr>
<td></td>
<td>More subjective words</td>
<td>More subjective words</td>
</tr>
</tbody>
</table>

### Table 1: IMDb biographies by race and gender.

<table>
<thead>
<tr>
<th>Race</th>
<th>Men</th>
<th>Women</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>African American</td>
<td>72</td>
<td>38</td>
<td>110</td>
</tr>
<tr>
<td>Caucasian</td>
<td>91</td>
<td>94</td>
<td>185</td>
</tr>
<tr>
<td>Total</td>
<td>163</td>
<td>132</td>
<td>295</td>
</tr>
</tbody>
</table>

---

10http://en.wikipedia.org/wiki/AFI’s_100_Years...100_Stars
11http://en.wikipedia.org/wiki/List_of_Best_Actor_winners_by_age
12http://en.wikipedia.org/wiki/List_of_Best_Actress_winners_by_age
Actor  | First sentence                                                                 | Prop. Adjs | Prop. Subj. |
-------|---------------------------------------------------------------------------------|------------|-------------|
Paul Newman | Screen legend, **superstar**, and the man with the most **famous** blue eyes in movie history, Paul Leonard Newman was born in January 1925, in Cleveland, Ohio, the second son of Theresa (Fetsko) and Arthur Sigmund Newman. | 3/36       | 2/36        |
John Amos | A native of New Jersey and son of a mechanic, African-American John Amos has relied on his **imposing** build, eruptive nature and **strong, forceful** looks to obtain acting jobs, and a **serious desire** for **better** roles to earn a satisfying place in the annals of film and TV. | 8/48       | 7/48        |
Amy Adams | Amy Lou Adams was born in Italy, to American parents Kathryn (Hicken) and Richard Kent Adams, while her father was a U.S. serviceman. | 1/23       | 0/23        |
Margaret Avery | **Slender, attractive** actress Margaret Avery, **spellbinding** in her role of Shug in Steven Spielberg’s The Color Purple (1985), is certainly no “one-hit wonder.” | 3/23       | 3/23        |

Table 2: First sentence for each of four sample biographies.

Descriptive Statistics

Since IMDb offers no information on the number or identity of user edits, the only available measure of effort made to document an actor’s life and career is biography length. Figure 3 depicts the distribution of biography length, by racial/gender group. As expected, the distribution is skewed to the right, with several outliers. The Kruskal-Wallis rank sum test (McKnight and Najab 2010), the non-parametric equivalent of ANOVA, detects no significance differences between the four groups, although biography length among African Americans varies more than within the groups of Caucasian Americans. For instance, the longest biography is that of Michael Jackson (16,836 words), and the shortest (30 words) describes Rusty Cundie. In the datasets of opening sentences (first five, first), the median number of words is 108 and 24, respectively. As illustrated in Figure 4, the opening sentences of biographies describing white men tend to be longer, as compared to the other groups.

Analysis

As mentioned, the LCM Manual emphasizes the differentiation of predicates involving adjectives versus verbs. Our automated analysis involves POS tagging and not full syntactic parsing. Nonetheless, if we observe salient differences between social groups with respect to the proportion of words used in biographies that are adjectives versus verbs, this would be a very good indication that the target individu-
als’ characteristics correlate to biographers’ tendencies to describe how they are (i.e., greater use of adjectives) versus what they have done (i.e., greater use of verbs). Likewise, frequent use of subjective or evaluative words would indicate a tendency to abstract away from concrete observations concerning the target individual.

Table 4 shows the proportion of adjectives and verbs used in the three datasets, broken out by the four social groups. We again use Kruskal-Wallis to test these differences. Post-hoc pairwise Wilcoxon rank sum tests using the Bonferroni correction confirm that on full texts, only the white men differ from the other groups. (In Tables 4 through 6, a bolded score indicates a group that differs from all others with p < 0.01). On the first five sentences of biographies, again, the post-hoc tests reveal that black men are described with fewer adjectives as compared to other groups. In the opening sentences of biographies, white men and black women are described with fewer verbs as compared to the other groups. In short, we begin to observe a pattern here: white actors tend to be described in a more abstract manner, with relatively greater use of adjectives and fewer verbs.

Table 5 displays the median proportion of words (any POS) that are strongly or weakly subjective. In full texts, white men are again shown to be described more abstractly, with a greater proportion of subjective words, as compared to the three other groups. In opening sentences, white men are described more abstractly than are black men or white women, but the post-hoc test reveals no difference between white men and black women.

As shown in Table 6, in the full texts, there are significant differences with respect to the proportion of adjectives used in a biography that are strongly subjective (i.e., carry sentiment regardless of context). The post-hoc test reveals that white men are described with more strongly subjective adjectives as compared to African Americans of both genders. In the opening sentences of biographies, the post-hoc test again reveals that white men and black women are described with more strongly subjective adjectives as compared to black men or white women.

Table 6: Median prop. of adj's that are strongly subjective.

![Figure 4: Distribution of words in first sentence.](image)

Table 5: Median prop. of subjective words (any POS).

<table>
<thead>
<tr>
<th></th>
<th>Full texts</th>
<th>First five</th>
<th>First sent.</th>
</tr>
</thead>
<tbody>
<tr>
<td>White men</td>
<td>0.03854</td>
<td>0.02597</td>
<td>0.02000</td>
</tr>
<tr>
<td>Black men</td>
<td>0.03401</td>
<td>0.03262</td>
<td>0</td>
</tr>
<tr>
<td>White women</td>
<td>0.03589</td>
<td>0.02405</td>
<td>0</td>
</tr>
<tr>
<td>Black women</td>
<td>0.03229</td>
<td>0.02692</td>
<td>0</td>
</tr>
<tr>
<td>Chi-square</td>
<td>12.1318</td>
<td>3.6413</td>
<td>7.8169</td>
</tr>
<tr>
<td>p-value</td>
<td>&lt;0.01</td>
<td>0.3029</td>
<td>&lt;0.05</td>
</tr>
</tbody>
</table>

Table 7: Prop. of texts with at least one strongly subjective adj.

<table>
<thead>
<tr>
<th></th>
<th>Full texts</th>
<th>First five</th>
<th>First sent.</th>
</tr>
</thead>
<tbody>
<tr>
<td>White men</td>
<td>0.945</td>
<td>0.5274</td>
<td>0.2418</td>
</tr>
<tr>
<td>Black men</td>
<td>0.750</td>
<td>0.361</td>
<td>0.1111</td>
</tr>
<tr>
<td>White women</td>
<td>0.904</td>
<td>0.394</td>
<td>0.0957</td>
</tr>
<tr>
<td>Black women</td>
<td>0.816</td>
<td>0.395</td>
<td>0.1579</td>
</tr>
<tr>
<td>Chi-square</td>
<td>219.6</td>
<td>39.02</td>
<td>186.92</td>
</tr>
<tr>
<td>p-value</td>
<td>&lt;0.0001</td>
<td>0.0001</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>

Finally, we use logistic regression to model the log odds of a biography containing a subjective adjective, based on the race and gender of the target individual (i.e., actor). More specifically, we fit a logit model as follows:

\[
\ln \frac{P(r_{yi} = 1)}{P(r_{yi} = 0)} = \beta_0 + \beta_1 \cdot Race_i + \beta_2 \cdot Gender_i + \beta_3 \cdot RG_i
\]

where \(\beta_0\) is the intercept, and the response, \(y_{yi}\), takes the value of 0 if the text is contains little abstract language (i.e., contains no subjective adjective) and 1 if the text is relatively more abstract (i.e., contains at least one strongly subjective adjective). Race is a binary variable where 1 indicates African American; likewise, Gender is 1 for actresses and 0 for actors. We also include an interaction term.

Table 8 shows the estimated model for each of the three datasets along with the p-value for the appropriate test of significance for each coefficient. In addition, the last row of
Across all three datasets, there is a significant main effect for race on the use of subjective adjectives, with African Americans’ biographies being less likely to contain abstract language. For the datasets consisting of the biographies’ first sentence(s), there is also a significant effect on gender: women are less likely to be described in abstract terms. Expected, we find the most subjective adjectives in the opening sentences describing white men. This is also the most diverse list; we find a total of 15 positive and five negative words. In contrast, in the first sentences of biographies of African American actresses, we find no subjective adjectives carrying negative sentiment.

We can also observe how IMDb biographers use subjective adjectives in the opening sentences. These words are used to describe a person’s appearance (e.g., “handsome,” “stunning”), disposition (“shrewd,” “charismatic”) and professional abilities (“talented,” “flexible”). Finally, it appears to be the case that biographers use negative adjectives primarily in the opening sentences of male actors. In general, these qualitative observations support the findings of the quantitative analyses: the race and gender of the biography subject correlate to the extent to which IMDb biographers’ use abstract and evaluative language in their descriptions.

Discussion

There is a growing literature surrounding the language of social media. Some scholars have documented the linguistic style of particular media, in order to better understand how users exploit and expand their communication affordances (e.g., (Hu, Talamadupula, and Kambhampati 2013))
or when and why people might change their linguistic style (Danescu-Niculescu-Mizil et al. 2013; Michael and Otterbacher 2014). In contrast, others have studied the correlation between participants’ linguistic patterns and their offline demographics and identities (e.g., (Pennacchioi and Popescu 2011; Fink, Kopecky, and Morawski 2012; Park et al. 2013). Both areas of research attempt to learn from the variations in language patterns to better understand people and their interactions via social media. However, previous work has not considered the possibility of linguistic biases in social platforms.

We framed and motivated our study of linguistic biases in social media in terms of the Linguistic Category Model, and the two most frequently studied biases, the Linguistic Expectancy Bias and the Linguistic Intergroup Bias. This body of research has already demonstrated that linguistic biases are commonplace in face-to-face interactions. In addition, they have demonstrated the effect that such biases have on how message recipients interpret the information conveyed. In short, linguistic bias tends to be very subtle, goes unnoticed much of the time, and yet it plays a salient role in the persistence of social stereotypes in society.

In our analyses, white men actors were consistently described with more abstract, interpretive language as compared to African American actors, and actresses of both races. Without a doubt, IMDb readers would notice the use of bluntly offensive or derogatory language in the biographies of actors and actresses. However, they may not consciously pick up on the subtle differences in the level of abstractness in the language used in the biographies. Because abstract descriptions of people tend to be much more powerful as compared to concrete descriptions, there is cause for concern. When presented with abstract descriptions, message recipients perceive the information conveyed as being more stable over time (Wigboldus, Spears, and Šemin 2005). In fact, in experimental settings, counter-stereotypical people can be made more likeable, being evaluated more positively by study participants, when they are described in a more abstract manner (Rubin, Paolini, and Crisp 2013).

We believe that the LEB and/or the LIB could take place at IMDb. LEB might occur when an IMDb participant attempts to describe an actor or actress, who somehow violates her expectations. When something or someone is unexpected, LEB predicts that the biographer would tend to focus on more concrete descriptions, avoiding the use of more abstract language that is subjective or inferential. Again, the reasons put forward thus far in the research indicate that the basis for this bias is cognitive (Karpinski and von Hippel 1996): unexpected information, which does not align with our existing prototypes, is more difficult for us to process.

For instance, a participant might not expect that an African American woman, Oprah Winfrey, who grew up in an underprivileged environment, could grow up to become one of the most famous, wealthy women in America. Winfrey’s IMDb biography begins with the following, quite concrete statements:

Born Orpah Gail Winfrey in Kosciusko, Mississippi, United States. Orpah was born to mother, former maid Vernita Lee, and father, war veteran Vermon Winfrey.

We can contrast the above to the opening lines of the biography of Arnold Schwarzenegger, a white male actor who also had a humble upbringing. As can be seen, the language used is more abstract, characterizing Schwarzenegger in a more generalizable way:

With an almost unpronounceable surname and a thick Austrian accent, who would have ever believed that a brash, quick talking bodybuilder from a small European village would become one of Hollywood’s biggest stars, marry into the prestigious Kennedy family, amass a fortune via shrewd investments and one day be the Governor of California!?

The LIB might occur because of the relationship between the IMDb participant and the target individual (i.e., the in-group or out-group status with respect to social attributes such as race and gender). If participation at IMDb is indeed male-dominated, as suggested by previous studies (e.g., (Hemphill and Otterbacher 2012)), then our results are expected. In other words, LIB predicts that white men biographers, the majority, are more likely to describe the achievements of white men actors in an abstract manner, and those of other social groups in a more concrete manner. In this way, they imply that their own group behaves in a positive way, and that achievement by other social groups are not usual (i.e., are not broadly generalizable) (Guerin 1994).

Limitations

Our study’s findings should be interpreted in light of its main limitation, which it shares with all observational studies of social media behavior: we report correlations but not causal relationships between the variables studied. More specifically, we were not able to control for confounding variables that might affect an IMDb biographer’s attitudes toward the actors and actresses in our dataset, as well as his or her writing style. Factors such as one’s age, ethnicity, or gender are very likely to affect the manner in which he or she writes (Labov 1990).

At IMDb, participation in the creation and editing of biographies is anonymous. In addition, the biographies are a collaborative effort. Only in a controlled experimental setting, where we could manipulate the social identities of the biographers and the target individuals, could we venture to say whether LIB, LEB or some other factor is the cause of the systematic differences in linguistic patterns that we have documented.

Nonetheless, the present work takes a very important first step toward developing methods to detect linguistic biases in social media descriptions of people. We have motivated the need for such studies, demonstrating how linguistic biases might manifest themselves through the use of abstract, subjective language. Finally, we have provided initial evidence that linguistic biases are, as expected, prevalent in social media just as they are in our offline interactions.
Conclusions

The current study laid the groundwork for a deeper analysis of linguistic biases in social media. Just as mass media researchers have cautioned that the presence of linguistic biases in messages can prime dominant stereotypes (e.g., in crime-related news and the variations in the descriptions of perpetrators as a function of race (Gorham 2006)), we would argue for further studies that gauge the extent to which LIB and LEB permeate our descriptions of people via social media, and with which consequences.

Studying linguistic biases in social media communication, as compared to offline interactions, adds a additional layer of complexity to these phenomena. We would not be at all surprised to observe differences in the types of biases, as well as in the frequency of their occurrences, were we to compare similar content and tasks (e.g., collaborative production of biographies of famous people) across social media platforms (e.g., Wikipedia vs. IMDb). Because the communicative context appears to mitigate the occurrence of bias (especially in the case of LIB (Maass et al. 1989)), future research must consider the role of factors such as: the specific communication affordances provided by a given platform, its participant base, and the social norms surrounding its use (e.g., the degree to which participants tend to self-disclose, whether or not participants maintain more friendly or formal relationships). In short, both the technical features and social cues surrounding the collaborative writing process likely correlate to the type of linguistic biases we observe in descriptions of people, as well as how often they occur. Only by achieving a better understanding of the relationships between such factors can we gauge whether or not the crowdsourcing of knowledge production also leads to the crowdsourcing of social stereotypes.

References


Labov, W. 1990. The Intersection of Sex and Social Class in the


