A quantitative analysis of gender differences in movies using psycholinguistic normatives

Anil Ramakrishna[1], Nikolaos Malandrakis[1], Elizabeth Staruk[1], Shrikanth Narayanan[1, 2]
[1]Department of Computer Science,
[2]Department of Electrical Engineering,
University of Southern California, Los Angeles, USA
{akramakr,malandra,staruk}@usc.edu, shri@sipi.usc.edu

Abstract

Direct content analysis reveals important details about movies including those of gender representations and potential biases. We investigate the differences between male and female character depictions in movies, based on patterns of language used. Specifically, we use an automatically generated lexicon of linguistic norms characterizing gender ladenness. We use multivariate analysis to investigate gender depictions and correlate them with elements of movie production. The proposed metric differentiates between male and female utterances and exhibits some interesting interactions with movie genres and the screenplay writer gender.

1 Introduction

Gender has been an important research topic in the social sciences, with studies conducted on the effect of gender on various aspects of human perception and expression (Benshoff and Griffin, 2011) as well as investigations of the societal (Behm-Morawitz and Mastro, 2008) and career implications of gender and possible underlying biases. Previous studies report significant implications of gender on career progress in medicine (Sidhu et al., 2009), information technology (Cohoon and Aspray, 2006), politics (Niven, 2006) and show-business (Smith, 2010).

In this paper we investigate the depictions of the genders in feature films, through the analysis of their respective dialogues. The differences in depiction are a contentious subject, since aspects of these can be viewed as the result of stereotyping or gender bias, with the relative presence of women being a well investigated subject (Bielby and Bielby, 1996; Lincoln and Allen, 2004). We are interested in the existing gender depictions, regardless of relative frequencies, as well as any factors that may affect them. While popular tools such as the Bechdel test provide a test for detecting female presence in the movies, we hope to identify more subtle forms of gender differences across character gender from the dialogues. Our aim is to devise a non-binary metric that can be used to compare or rank movies, characters and perhaps individual utterances.

To analyze the dialogues we propose using a metric of language gender ladenness, a number representing a normative rating of the "perceived feminine or masculine association" (Paivio et al., 1968) of language. The metric, as originally defined, is meant to provide an indication of gender-specificity of individual words, with extreme values assigned to highly stereotypical concepts. Generating this rating for male and female character dialogues and comparing the character gender with this rating of "language gender" should allow us to observe stereotypical behavior.

Word based ratings such as the gender ladenness are referred to as linguistic norms (or psycholinguistic norms when corresponding to psychological constructs) and are popular in cognitive psychology (Clark and Paivio, 2004) and some computational disciplines, such as sentiment analysis (Nielsen, 2011) and opinion mining. To utilize gender ladenness, we follow an approach similar to simple sentiment analysis, with word-level norms automatically generated based on a small starting set of manually annotated norms and sentence (and higher) level norms estimated through word-level norm statistics. The resulting algorithm allows us to estimate gender ladenness at any arbitrary granularity.

We use these ratings of dialogue language to quantify the depictions of male and female characters and attempt to relate the observed gender ladenness with objective factors.

In section 2 and 3 we describe the data corpus

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 1996–2001,
Lisbon, Portugal, 17-21 September 2015. ©2015 Association for Computational Linguistics.

Word	*Norm*
Infantry	-0.97
Truck	-0.73
Dictator	-0.56
Strider	-0.36
United	-0.18
Volunteerism	0.04
Hygiene	0.22
Candle	0.45
Radiant	0.66
Bride	0.84
Gorgeous	0.96

Table 1: Sample word norms($\in [-1, 1]$); -1: Most masculine and $+1$: Most feminine.

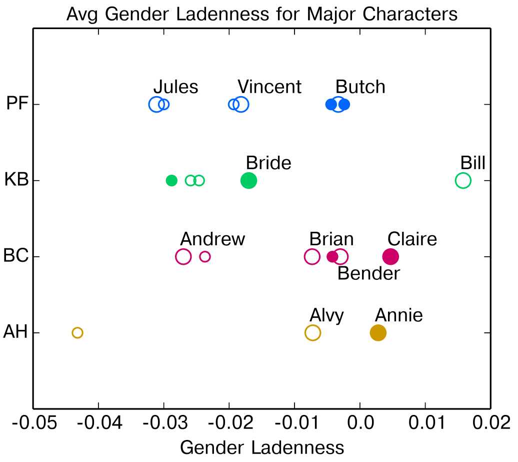


Figure 1: Average Gender Ladenness for a few sample movies, marker size proportional to number of utterances. Filled markers: Female characters, Hollow markers: Male characters; PF: Pulp Fiction, KB: Kill Bill, BC: Breakfast Club, AH: Annie Hall.

and the feature extraction process respectively. We detail the experimental procedure in section 4 and analysis in section 5 and conclude with future extensions in section 6.

2 Estimating Gender Ladenness

Gender Ladenness, as defined in (Clark and Paivio, 2004) represents the degree of perceived "feminine or masculine association" on a numerical scale ranging from very masculine to very feminine. It is important to note that there was no restriction to what "association" may mean: while it is reasonable to assume that associations of the form "A is B" or "A has B" would dominate annotator perception, that does not preclude other forms of association. Because of that, referring to the norms as indicators of how masculine or feminine the words are is not entirely accurate, though it is a reasonable approximation. The original ratings were re-scaled to $[-1, 1]$ for our purposes, with lower values indicating a masculine association and high values indicating a feminine association. Some sample words, utterances and their corresponding ratings are presented in Table 1 and Table 3. Figure 1 shows the average gender ladenness across all utterances for the major characters of a few movies. The annotations as a whole are reflective of stereotypical views of gender roles, e.g., words related to war and violence have a strong masculine association, whereas words related to family or positive emotions carry strong feminine associations.

The manual annotations from (Clark and Paivio, 2004) contain ratings for only 925 words, which are not enough to provide sufficient coverage. Therefore we use a lexicon expansion method, inspired by the work of (Malandrakis et al., 2013) to estimate the gender ladenness $\hat{g}(w_i)$ of word w_i using the semantic similarities $s()$ between w_i and reference words or concepts c_j, as

$$\hat{g}(w_i) = \theta_0 + \sum_{j=1}^{N} \theta_j s(w_i, c_j), \qquad (1)$$

where the terms θ_i are trained model parameters.

Given a manually annotated lexicon and a set of reference words, this equation can be used to create a linear system. Solving the system via Least Squares Estimation (LSE) gives us the parameters θ and an equation that can be used to generate gender ladenness for any new set of words.

Gender ladenness for larger lexical units is generated via simple statistics, as the average of word gender ladenness over all content words (adjectives, nouns, verbs and adverbs).

3 Data

Our primary data source is the Movie DiC corpus (Banchs, 2012) which includes 619 movie scripts parsed from The Internet Movie Script Database (IMSDb, 2015). The xml formatted scripts contain transcripts with speaker information as well as some structural information. Additional metadata for each movie were collected from the Internet Movie Database (IMDb, 2015).

Since our goal was to analyze gender depictions, we had to annotate each script utterance with a gender label. The process was complicated by inconsistencies between the information contained in the IMDb and Movie DiC corpora, like mismatched names, particularly for minor characters. Initially the script character names were cleaned using simple heuristics, such as the removal of all instances of the possessive "'s". The IMDb api (IMDbPY, 2015) was used to recover candidate movies matching the script movie name and, in the case of multiple candidates, the best candidate was selected based on the number of character names matching the script. Character names were compared using the Jaro-Winkler distance (Winkler, 1990). Having achieved a one to one mapping between IMDb and Movie DiC, we assigned a gender label to each matched character, using the gender predictor (NamSor Applied Onomastics, 2015). To make these predictions, we first use the name of the corresponding actor portraying that role; if there was no character match, we use the name of the character. Finally, we calculate a *confidence* score of our gender assignment per utterance for each movie, equal to the percentage of utterances with perfectly matched character name and a high confidence by the gender predictor. For the movies for which the confidence scores are not satisfactory, we manually match the script characters with IMDb's characters and annotate genders. In our experiments, we did this manual annotation with roughly 75 movies.

Having a mapping of scripts to IMDb entries, we collected more information about the movie such as the list of genres it belongs to and the members of the production team (producers, scriptwriters, directors), and followed a similar process as described above to assign genders to all persons. While movies may be created by multiple scriptwriters and directors, we retain only the first name, the primary credit, in each category. We removed infrequent genres and movies which belonged only to the removed genres. We also filtered out utterances with missing or incorrect character information and the utterances corresponding to characters for which the gender predictor fails to make confident predictions.

Movies with missing fields were also removed, leaving us with a total of 568 movies after the aforementioned pre-processing steps. Table 2 lists some descriptive statistics of the processed movie corpus. At least in terms of raw frequencies, the gender ratio is clearly skewed towards male, particularly in the case of directors and with the exception of casting directors.

Property Name	*Female*	*Male*	*Total*
Movie Utterances	107372	256274	363646
Producers	746	2974	3720
Directors	33	572	605
Assistant Directors	846	2739	3585
Screenplay Writers	130	1217	1347
Casting Directors	548	195	743

Table 2: Movie Dataset statistics

Utterance	*Avg. GL*
Flowers for the Diva.	0.77
Yeah, what a woman.	0.47
Got the house to yourselves?	-0.01
What about the crew?	-0.51
Yeah? You and what army?	-0.85

Table 3: Average gender ladenness for sample utterances from the dataset

The norm generating equation (1) requires a semantic similarity estimate $s()$, which for the purposes of this paper is the cosine of context vectors generated over a large corpus of raw web text. The corpus was created by posing a query to the Yahoo! search engine for every word in the English version of the aspell spell-checker and collecting the top 500 result previews. Each preview is composed of a title and a sample of the content, each being a single sentence. Overall the collected corpus contains approximately 117 million sentences.

4 Experimental Procedure

The descriptive feature in this method is gender ladenness, so we extracted an estimate for each utterance of every movie. Initially, all utterances were part-of-speech tagged and non-content words were removed. Then, word-level gender landenness norms were generated for every remaining word.

To generate word-level norms, we used equation (1) with the intermediate seed words w_i being the top 10000 most frequent words in our corpus of web text with length longer that 3 characters. For each word in our corpus, we generated a binary weighted context vector (of window size 1) of size $\sim$ 125000. Then, for each word of interest we calculated a 10000 place similarity vector, containing the cosine similarity scores between the context vector of said word and the context vectors of the 10000 intermediate seeds. Using the training set we generated a $K \times 10000$ matrix of similarities to the seed words and applied dimensionality reduction via Principal Component Analysis (PCA), keeping the first $N = 300$ components. These transformed similarities became the similarity terms $s()$ of equation (1) and were used to train the model. For any word in the scripts, a 10000 place similarity vector is generated and transformed using the pre-calculated PCA matrix, then equation (1) is used to create the gender ladenness estimate.

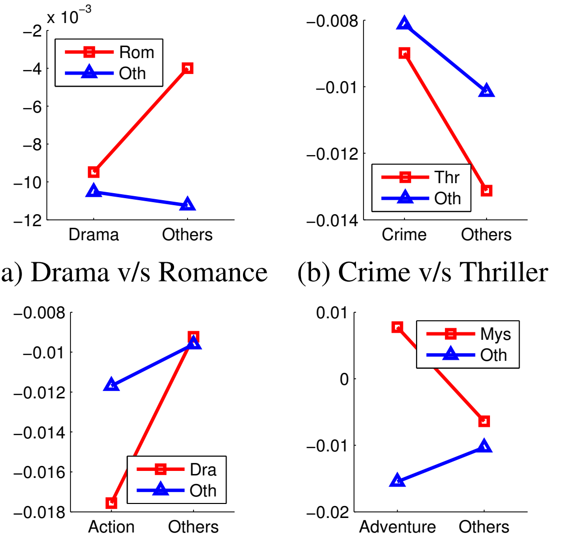
Ratings were generated at the utterance level, and collective ratings (per character, gender or movie) were calculated as utterance rating averages.

5 Results

To evaluate the word norm generation algorithm, we performed a 10-fold cross-validation experiment on the 925 manually annotated norms in (Paivio et al., 1968). The generated norms were evaluated against the ground truth and the method achieved a 0.801 Pearson correlation to the ground truth. While there is no comparable result in literature, the resulting performance appears sufficiently high.

We first investigated the overall gender ladenness of movies, represented as the average of all utterance level scores, with respect to the genre(s) the movie belongs to. The independent variables for this analysis were nine binary indicator variables, one for each of the most frequent genre labels in our movie corpus, with values of zero if the movie does not belong to the specific genre and one if it does. The particular representation of genres as separate variables was chosen because each movie can belong to multiple genres. Interaction terms were included. Running n-way ANOVA with the aggregate gender ladenness across both character genders as the dependent variable revealed significant differences between genres, with *Action* movies leaning towards the masculine (p = 0.013) compared to *Non-Action* movies, a not surprising result.

A few significant interactions between genres are shown in figure 2. Fig. 2a indicates that among non-drama movies, romantic movies tend to include more feminine language compared to non-romantic movies. However, if a movie belongs to the genre drama, its mean gender ladenness scores remain fairly constant, irrespective of its other genres. Similar interpretations can be drawn from the other plots in figure 2.



(a) Drama v/s Romance (b) Crime v/s Thriller

(c) Action v/s Drama (d) Adventure v/s Mystery

Figure 2: Interactions between genres

To analyze the effect of character gender on the gender ladenness scores, we next ran ANOVA with the character gender and the movie writer's gender as additional independent variables. The dependent variable in this case was the aggregate gender ladenness score across all utterances for male and female characters, so two scores per movie. The interaction of character gender and movie genre is shown in figure 3. The scores of male and female characters are correlated, which can be attributed to the underlying concepts in the utterances from these movies. The difference between genders is significant ($p = 0.034$), with male characters consistently using significantly more masculine language than their female counterparts, a finding that lends some credence to the metric used. Looking at the binary genre variables revealed that

Action movies contained significantly more masculine language than *Non-Action* movies (p $< 10^{-5}$) and the same holds for *Crime* movies (p $< 10^{-5}$). Conversely, *Romantic* movies leaned towards the more feminine language than non-*Romantic* movies ($p < 10^{-5}$) and similarly for *Comedy* movies compared to non-*Comedy* movies ($p = 0.02$). The male - female character gender

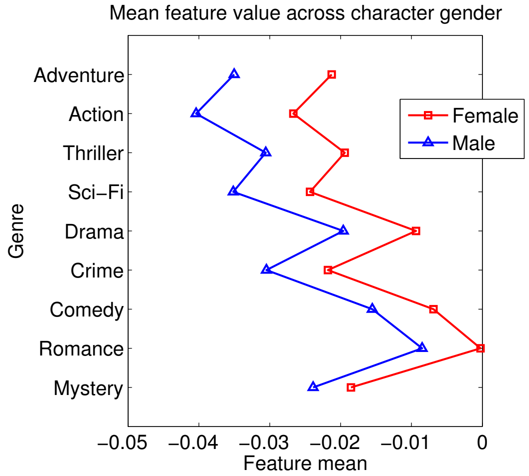


Figure 3: Interaction of movie genre with character gender

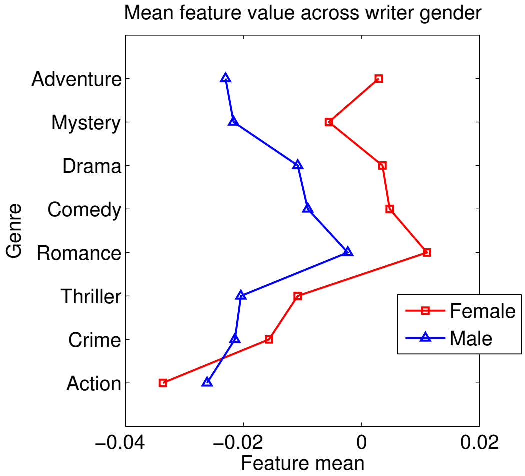


Figure 4: Interaction of screenplay writer's gender with genre

ladenness distance however is not affected in any significant way by the genre.

We include only the screenplay writer's gender in our analysis, though both the directors and screenplay writers influence the dialog lines (utterances), since the writers are more likely to directly influence the actual language used. In addition, the very small number of female directors in the data, as seen in table 2, leads to a violation of ANOVA's homoscedasticity assumption. Though the writer gender itself was not a significant factor, the interaction of writer's gender with the *Action* genre was significant ($p = 0.005$). The plot illustrating this interaction is shown in figure 4. It appears that female script writers write more masculine utterances compared to their male colleagues, at least for *Action* movies. We also investigated interactions between the writer and character gender, but none proved significant.

6 Conclusions and Future Work

We used regression to extrapolate manually annotated psycholinguistic normatives to movie utterances and investigated the use of these metrics to describe gender depictions. The metric proved successful, showing significant differences between the genders and predictable patterns with respect to movie genres.

Future work will include the use of further metrics, with those describing emotions being the first candidates. We also intend to collect more movie and character level metadata to be used in analysis. Finally, it is worth remembering that language provides only a partial description of depicted characters, so we should aim to augment with aural/visual information.

7 Acknowledgements

The authors gratefully acknowledge support from NSF, Geena Davis Institute on Gender in Media and Google.org.

References

Rafael E Banchs. 2012. Movie-Dic: a movie dialogue corpus for research and development. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pages 203–207. Association for Computational Linguistics.

Elizabeth Behm-Morawitz and Dana E Mastro. 2008. Mean girls? the influence of gender portrayals in teen movies on emerging adults' gender-based attitudes and beliefs. *Journalism & Mass Communication Quarterly*, 85(1):131–146.

Harry M Benshoff and Sean Griffin. 2011. *America on film: Representing race, class, gender, and sexuality at the movies*. John Wiley & Sons.

Denise D Bielby and William T Bielby. 1996. Women and men in film gender inequality among writers in a culture industry. *Gender & Society*, 10(3):248–270.

James M Clark and Allan Paivio. 2004. Extensions of the Paivio, Yuille, and Madigan (1968) norms. *Behavior Research Methods, Instruments, & Computers*, 36(3):371–383.

J McGrath Cohoon and William Aspray. 2006. *Women and information technology: Research on underrepresentation*, volume 1. The MIT Press.

IMDb. 2015. Internet movie database. [Online; accessed 10-June-2015].

IMDbPY. 2015. [Online; accessed 10-June-2015].

IMSDb. 2015. Internet movie script database. [Online; accessed 10-June-2015].

Anne E Lincoln and Michael Patrick Allen. 2004. Double jeopardy in hollywood: Age and gender in the careers of film actors, 1926–1999. *Sociological Forum*, 19(4):611–631.

Nikolaos Malandrakis, Alexandros Potamianos, Elias Iosif, and Shrikanth Narayanan. 2013. Distributional semantic models for affective text analysis. *Audio, Speech, and Language Processing, IEEE Transactions on*, 21(11):2379–2392, Nov.

NamSor Applied Onomastics. 2015. NamSor gender API. [Online; accessed 10-June-2015].

Finn Årup Nielsen. 2011. A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. *arXiv preprint arXiv:1103.2903*.

David Niven. 2006. Throwing your hat out of the ring: Negative recruitment and the gender imbalance in state legislative candidacy. *Politics & Gender*, 2(04):473–489.

Allan Paivio, John C Yuille, and Stephen A Madigan. 1968. Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of experimental psychology*, 76(1p2):1.

Reena Sidhu, Praveen Rajashekhar, Victoria L Lavin, Joanne Parry, James Attwood, Anita Holdcroft, and David S Sanders. 2009. The gender imbalance in academic medicine: a study of female authorship in the united kingdom. *Journal of the Royal Society of Medicine*, 102(8):337–342.

Stacy L Smith. 2010. Gender oppression in cinematic content? a look at females on-screen & behind-the-camera in top-grossing 2007 films. *Retrieved September*, 2:2010.

William E Winkler. 1990. String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage.