Gender-Distinguishing Features in Film Dialogue

Alexandra Schofield & Leo Mehr
Cornell University
Ithaca, NY 14850

Abstract

Film scripts provide a means of examining generalized western social perceptions of accepted human behavior. In particular, we focus on how dialogue in films describes gender, identifying linguistic and structural differences in speech for men and women and in same and different-gendered pairs. Using the Cornell Movie-Dialogs Corpus (Danescu-Niculescu-Mizil et al., 2012a), we identify significant linguistic and structural features of dialogue that differentiate genders in conversation and analyze how those effects relate to existing literature on gender in film.

1 Introduction

Film characterizations often rely on archetypes as shorthand to conserve narrative space. This effect comes out strongly when examining gender representations in films: assumptions about stereotypical gender roles can help establish expectations for characters and tension. It is also worth examining whether the gendered behavior in film reflects known language differences across gender lines, such as women's tendency towards speaking less or more politely (Lakoff, 1973), or the phenomenon of "troubles talk," a ritual in which women build relationships through talking about frustrating experiences or problems in their lives (Jefferson, 1988) in contrast to a more male process of using language primarily as a means of retaining status and attention (Tannen, 1991). We look at a large sample of scripts from well-known films to try to better understand how features of conversation vary with character gender.

We begin by examining utterances made by individual characters across a film, focusing on the classification task of identifying whether a speaker is male or female. We hypothesize that in film, speech between the two gender classes differs significantly. We isolate interesting lexical and structural features from the language models associated with male and female speech, subdividing to examine particular film genres to evaluate whether features are systematically different across all genres or whether distinguishing features differ on a per-genre basis.

We then focus on the text of conversations between two characters to identify whether the two speakers are both male, both female, or of opposite genders. One belief about gendered conversation expressed in films is that women and men act fundamentally differently around each other than around people of the same gender, due partly to differences in the function of speech as perceived by men and women (Tannen, 1991). We look into features that explore the hypothesis that there are significant differences in how men and women speak to each other that are not accounted for merely by the combination of a male and a female language model, and find distinguishing features in each of these three classes of language. Finally, we look at whether these conversation features have predictive power on the duration of a relationship in a film.

2 Data Description

Our dataset comes from the Cornell Movie-Dialogs Corpus (Danescu-Niculescu-Mizil et al., 2012a), a collection of dialogues from 617 film scripts. Of the characters in the corpus, 3015 have pre-existing gender labels. We obtain another 1358 gender labels for the remaining characters by taking the top 1000 US baby names for boys and girls and treating any character whose first listed name is on only one of these two lists as having the respective gender

Proceedings of the Fifth Workshop on Computational Linguistics for Literature, NAACL-HLT 2016, pages 32–39,
San Diego, California, June 1, 2016.

of the list. Based on hand-verification of a sample of 100 these newly-added labels, we achieved 94% labeling accuracy, implying that the 4373 character labels have about 98% accuracy. In practice, many of the mislabeled names seem to be from characters named for their job title or last name, suggesting that these characters have fairly little contribution to the dialogue. We investigated using IMDb data as an additional resource but discovered that variations in character naming make this task complex.

Women are less prominent than men across all films, both possessing fewer roles (30% of all roles in major films in 2014) and a smaller proportion of lead roles within them (Lauzen, 2015). This observation is matched quite well in the Movie-Dialogs corpus, where after supplementing gender labels, only 33.1% of characters are female (previously, 32.0% of the original characters were female). In addition, we record 4676 unique relationships (judged by having one or more conversations) with known character genders. A chi-squared test to compare the expected distribution of gender pairs from our character set to the actual relationships shows that the characters are not intermingling independently of gender ($p < 10^{-5}$), with only 374 of the expected 509 relationships between women and 2225 interactions between men compared to the expected 2099.

Subdividing our data further, we find that certain film genres as represented in this dataset have disproportionate representation of certain gender pairs with respect to gender. Table 1 shows the significant differences within genders of actual vs. expected number of characters and relationships of each gender type. Though we hypothesized that the gender gap may have narrowed over time, we find the gender ratio fairly consistent across time in our corpus, as shown in Figure 1.

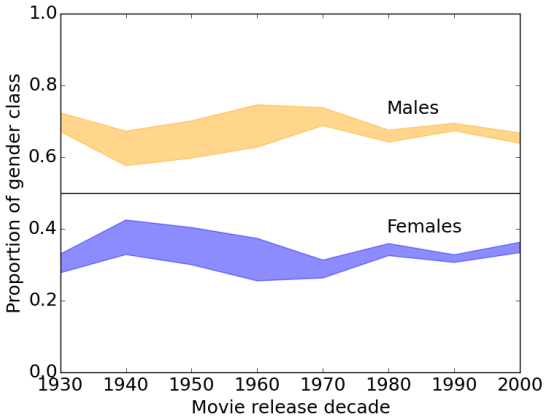


Figure 1: Proportion of character gender representation in movies, bucketed by decade, shaded by standard error.

Category	Key	Features
Lexical	LEX	unigrams, bigrams, trigrams
Vader Sentiment Scores	VADER	VADER scores for positive, negative, neutral, and composite value
Valence, Arousal, and Dominance	V/A/D	average scores across scored words
Structural	STR	average tokens per line, average token length, type to token ratio
Discourse	DIS	Δ average tokens per line, Δ average token length, Δ type to token ratio, unigram similarity

Table 2: List of feature groups. Δ indicates the absolute, unsigned difference between the text for each speaker. We discarded LEX features that arose fewer than 5 times.

3 Methods

3.1 Feature Engineering

Our text processing uses the Natural Language Toolkit (NLTK) (Bird et al., 2009). We use a simple tokenizer in our analysis that treats any sequence of alphanumerics as a word for our classifiers, splitting on punctuation and whitespace characters. We elect not to stem or remove stopwords, as non-contentful variation in language is important for our analysis.

Based on theory that women will have more hedging (Lakoff, 1973), we hypothesized that strength of sentiment or signals of arousal or dominance might also signal gender differences in convesation. We used sentiment labels from VADER (Hutto and Gilbert, 2014) and a list of 13,915 English words with scores describing valence, arousal, and dominance (Warriner et al., 2013). We group these features and several nonlexical discourse features into several primary groups, described in Table 2. We also experimented with part-of-speech labels using the Stanford POS tagger (Toutanova et al., 2003), but found they do not significantly influence results.

Genre	M	F		MM	FM	FF	
action	735	295	**	562	434	40	****
adventure	486	184	**	388	284	17	****
animation	82	34		68	41	5	
biography	156	63		128	80	13	
comedy	857	430		695	636	147	
crime	750	299	**	604	427	68	**
drama	1645	830		1278	1192	195	****
family	74	40		43	62	9	
fantasy	314	158		246	232	42	
history	95	42		80	46	5	**
horror	365	245	***	209	338	89	*
music	67	35		62	48	4	**
mystery	496	243		403	364	63	**
romance	660	372	*	463	566	119	*
sci-fi	502	205	*	381	321	27	****
thriller	1240	575		918	810	133	***
war	114	29	**	99	48	3	
western	79	40		66	51	12	

Table 1: Chi-squared test results on number of characters of each gender and number of gender relationship pairs given gender proportions. The character gender test is done in comparison to the 33% female baseline expectation for that number of characters, whereas the gender-pairs are with respect to the expected proportion of gender pairs were one to randomly draw two characters for each of the relationships observed. Only genres with more than 100 observed characters with assigned gender were included. Stars mark significance levels of p=0.05*, 0.01**, 0.001***, and 0.0001****.

We surveyed several types of simple classifiers in our prediction tasks: Gaussian and Multinomial Naive Bayes, and Logistic Regression. These implementations came from the **scikit-learn** Python library (Pedregosa et al., 2011).

3.2 Controlling Data

In comparing the language of males and females, we want to ensure that confounding factors do not result in significant results; the classification tasks should not yield better/worse results because of the structure of our dataset or the data we used to train/test. The first essential measure we take is to select equal numbers of males and females from each movie. Second, we only further select characters that have non-trivial amount of speech in the film. When selecting characters for single-speaker analysis, we use only those which had at least 3 conversations with other characters, 10 utterances, and 100 words spoken in total. This removes 45% of the characters from the original dataset. While the specific numbers are arbitrary, they were roughly selected after examining random character dialogs by hand. Third, we control for the language of a given movie or the style of its screenwriter(s) by using a leave-one-label-out split when running our classifiers.

Similarly for conversations, we control for each of the gender classes (male-male, female-male, and female-female), by including from each film the same number of conversations from each class. This results in a set of roughly 3500 conversations for consideration, a substantial subset of the original corpus but one with representation of a variety of dialogue lengths and less affected by the gender variation within particular films, to avoid classifying film content.

4 Experiments

4.1 Evaluating Individual Gender Features

We first examine the language differences in male and female utterances, selecting an equal number k_i of random male and female characters from each movie i. We then develop language models based upon the unigram, bigram, and trigram frequencies across all utterances from selected male characters versus female characters. As our focus is on usage of common words, we use raw term frequency instead of boolean features or TF-IDF weighting. While this does not fully control for the amount of speech of a

given gender, it does control for variation in gender ratios and conversation subjects within films and genres.

We analyze the interesting n-grams using the weighted log-odds ratio metric with an informative Dirichlet prior (Monroe et al., 2008), distinguishing the significant tokens based upon single-tailed z-scores. Notably, with a large vocabulary, it is expected that some terms will randomly have large z-scores. We therefore only highlight n-grams with z-scores of greater magnitude than what arose in 19 out of 20 tests of random reshufflings of the lines of dialogue between gender classes (equivalent to the 95% certainty level of what is significant). The important n-grams are displayed in Figure 2.

The findings here conform to findings we would expect, such as cursing as a male-favored practice (Cressman et al., 2009) and polite words like greetings and "please" as more favored by women (Holmes, 2013). Interesting as well is the predominance of references to women in men's speech and men in women's speech: "she" and "her" are strongly favored by male speakers, while "he" and "him" are strongly favored by female speakers ($p < 0.00001$). We also observe that in contrast to men's cursing, adverbial emphatics like "so", and "really" are favored by women, conforming to classic hypothesis about gendered language in the real world (Pennebaker et al., 2003; Lakoff, 1973).

4.2 Predicting Speaker Gender

Given only the words a character has spoken in conversations over the course of the movie, can we accurately predict the character gender?

As outlined in Controlling Data, we select characters equitably from each movie, each having spoken a significant amount during the movie. Using this method, we obtain 552 male and female characters each. We extract features from the all the lines spoken by each of these characters (as outlined in Feature Engineering), and train/test various scikit-learn built-in classifiers (as from Classifiers) in 10-fold cross-validation. As surveyed here, using a Logistic Regression classifier with different features, we obtain 72.2% classification accuracy (per feature accuracy outlined in Table 3). A multinomial Naive Bayes classifier performs slightly better, on which we applied the more appropriate leave-one-label-out cross-validation method to split training and test data, at **73.6%**.

Features	Accuracy±Std. Error
Baseline	$50.0 \pm 0.3\%$
STR	$55.2 \pm 2.1\%$
Unigrams	$67.4 \pm 1.7\%$
LEX	$\mathbf{71.7 \pm 1.9\%}$
LEX + STR	$\mathbf{72.0 \pm 1.9\%}$
LEX + STR + VADER	$\mathbf{72.2 \pm 1.2\%}$

Table 3: Performance of single-speaker gender classification. Bolded outcomes are those statistically insignificantly different from the best result (using a two-tailed z-test).

4.3 Evaluating Relationship Text

While the previous section demonstrates systemic differences in language between male and female speakers, an additional factor to consider is the conversation participants of each of these dialogues. We can hypothesize that, in addition to having different lexical content between men and women, movies also demonstrate significant content differences between pairs of interacting genders, such that the conversation patterns of men and women talking to each other have different content than same-gendered conversations.

We can examine this hypothesis by repeating the analysis performed on single characters throughout a film on individual conversations from films. We use the controlled dataset described in the Methods section, this time contrasting each class of gender pair: male-male, female-male, and female-female (MM, FM, and FF, respectively). We include the most significant words in each class in Table 4. As with the single-gender analysis, we see that men seem to speak about women with other men, and women about men with other women. We also note that several pronouns including "she" and "he" from before are actually considered statistically less probable in two-gendered conversations.

This is an interesting signal of men speaking differently around men than around women, which, in conjunction with the high log-odds ratio of "feel", "you", and "you love" favoring dual-gendered conversations, suggests that men and women are more likely to be talking about feelings and each other, while they are more likely to talk about experiences

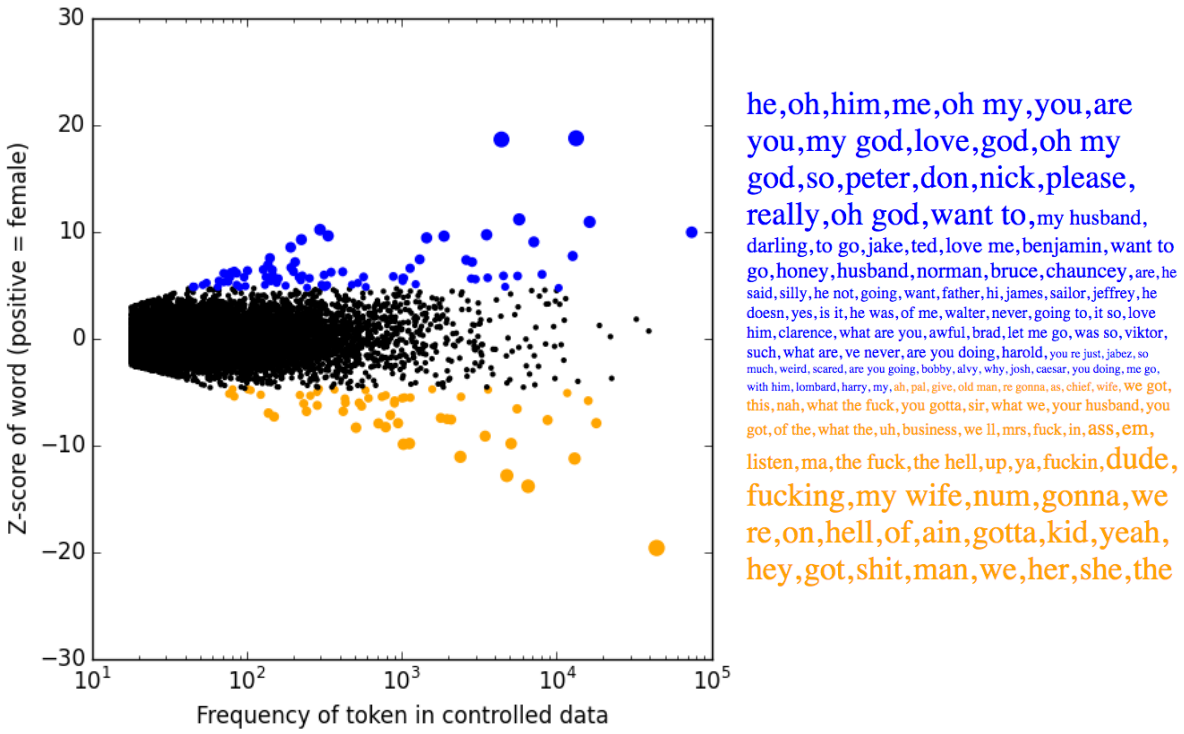


Figure 2: Tokens with significance plotted with respect to log-odds ratio. We ran 20 randomization trials and found that in those trials, the largest magnitude z-score we saw was 4.7. Blue labels at the top refer to female words above that significance magnitude, while orange labels at the bottom refer to words below that significance.

of the other-gendered people in their lives with their same-gendered friends. While this finding does not fully support that women and men are not friends in films, it does suggest the idea that men and women in films are typically interacting in a way distinct from men and women without consideration of context. It also contrasts with the typical understanding of sharing personal problems as a female practice (Tannen, 1991), as it seems that both men and women in films use words discussing feelings and people of the other gender.

4.4 Predicting Gender Pairs

In order to focus on the linguistic differences of the content of conversations between our gender pair classes instead of the success of per-character gender classifiers, we took as our additional classification task the problem of predicting the gender pair of the speakers in a conversation. This task is considerably more difficult than most, as conversations are often short and will include multiple speakers. We again use leave-one-label-out training to avoid learning dialogue cues from movies. While we can again attain better accuracy with a multinomial Naive Bayes classifier on LEX features, for our objective of simply demonstrating that features provide indication of gender differences, we are satisfied to use logistic regression to incorporate all features.

As Table 5 shows, the only features producing significant improvement over a random accuracy baseline of 33% are lexical, structural, and discourse features. While the fact that lexical content has distinguishing power is perhaps unsurprising, given the preceding analysis, it is somewhat more surprising that more simple structural and discourse features are also producing significant results.

While there no obvious significant structural differences, one can spot minor variation that seems to provide the slight improvement above random in our classification in Figure 3. We observe in Figure 3a

MM		FM		FF	
n-gram	z	n-gram	z	n-gram	z
her	8.2	feel	3.9	he s	9.0
she	7.7	you	3.5	he	7.2
the	7.0	you love	3.0	him	6.3
man	6.7	walk	2.8	he was	4.6
this	4.6	happy	2.8	dear	4.2
sir	4.3	tough	2.8	honey	4.0
you	−3.6	in my	−2.6	up	−3.4
honey	−3.8	every	−2.8	man	−3.8
him	−4.4	man	−3.1	her	−4.3
love	−4.8	she	−3.4	she	−4.5
he	−4.8	he	−4.2	mr	−4.5
hes	−4.9	her	−4.2	the	−5.2

Table 4: The six top words and z-scores correlated with the topic positively and negatively when comparing log-odds ratios for each gender class with respect to the other two. While a z-score of magnitude 2.8 has a significance of $p < 0.003$, the size of the considered vocabulary makes it unsurprising that several words have scores of this magnitude randomly; however, in twenty trials of randomization of the text between classes, only one z-scores emerged greater than magnitude 3.1. We therefore infer z-scores higher than 3.1 or lower than -3.1 are unlikely to be the consequence of random variation between classes.

that while utterance length is significantly higher for all-male than all-female conversations, two-gender conversations seem to behave more like all-female conversations on average. Figure 3b looks again at speaker utterances in combination with their imbalance between speakers, the "delta" average utterance length. Our comparison shows a significant difference between men talking to men and men talking to women. As delta utterance length here explicitly is described by average female utterance length minus average male utterance length, this demonstrates that women are speaking in shorter utterances than men in male-female conversations, in contrast to having longer utterances overall. Word length also is significantly shorter for women than men in single-gender conversations, but in this case, the two-gendered value appears to be just the interpolation of the two single-gender values, suggesting that word length is not decreased for male characters in two-gender conversation.

We also can see some interesting discourse features in Figure 3c. While looking at the data confirms that the average type-to-token ratio does not

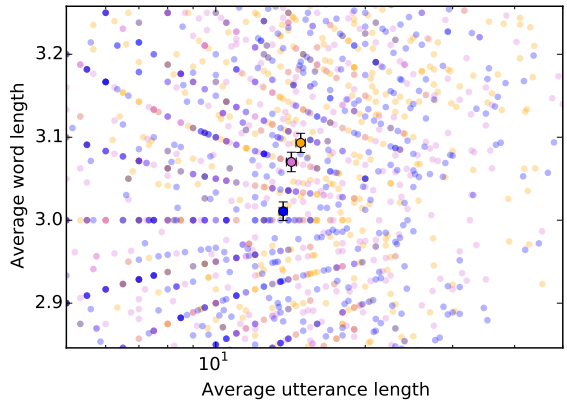
Features	Accuracy±Std. Error
LEX	**38±1%**
VADER	33±1%
V/A/D	35±1%
STR + DIS	**37±1%**
LEX + STR + DIS	**37±1%**
All but LEX	35±1%
All	**38±1%**

Table 5: Classifier results using logistic regression on the features from Table 2. Lexical features are sufficient to produce nonrandom classification, as well as structural and discourse features. Bolded text indicates a result better than random ($p < 0.05$).

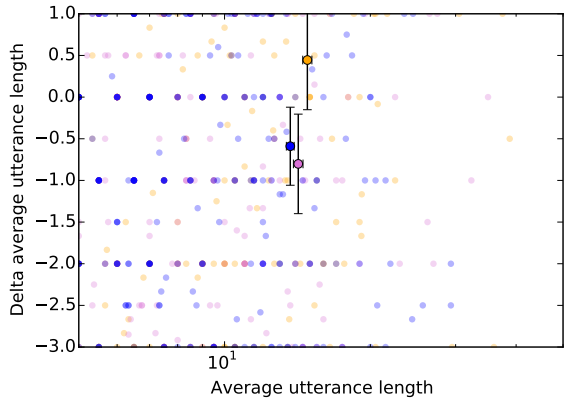
differ between our three conversation classes, we find that the type-token ratio difference is significantly higher for conversations between two genders, which suggests that two-gender conversations may have an increased probability of demonstrating one character as less articulate than another. Looking into the data, this slightly but insignificantly favors women having a higher type-to-token ratio than men, suggesting they use more unique words in their speech than do men in conversation. Finally, we note that conversations with women have significantly higher unigram similarity than men. This hints there may be some linguistic mirroring effect that women in film demonstrate more than men, which may relate to the hypothesis that women coordinate language more to build relationships (Danescu-Niculescu-Mizil et al., 2012b; Tannen, 1991).

4.5 Relationship Prediction

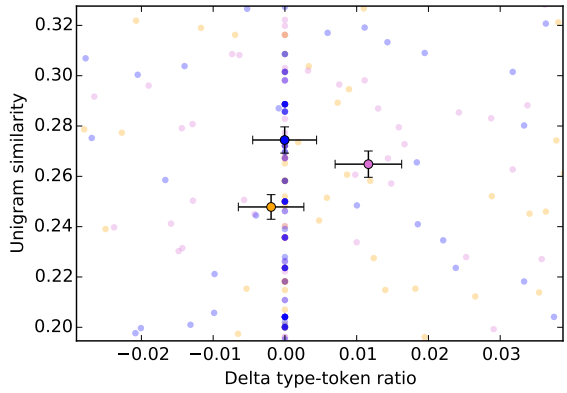
In addition to testing the prediction of genders in conversations and relationships, we attempted to use the same features to distinguish from a single conversation whether a relationship would be short (3 or fewer conversations) or long (more than 3 conversations). We tested on a dataset of conversations split evenly between gender pairs and between long and short relationships, using leave-one-label-out cross validation to test conversations from one relationship at a time. With a multinomial Naive Bayes classifier, we are able to achieve $60 \pm 2\%$ accuracy with a combination of n-gram features, gender labels, and structural and discourse features. Performing ablation with each feature set used, we find that results worsen by omitting either structural features ($54 \pm 2\%$) or n-gram features ($54 \pm 2\%$), but that omitting gender from the classification does



(a) Structural features.



(b) Utterance length.



(c) Discourse features.

Figure 3: Structural and discourse features plotted with respect to each other, focusing on the region of means (circled in black). Orange and blue refer to male-male and female-female conversations, while pink refers to two-gender conversations. Standard errors for both axes are plotted in each figure but are sometimes too small to distinguish.

not significantly impact the classification accuracy ($60 \pm 2\%$).

Some of this result is predictable from the limits of the data: controlling for the number of conversations in a relationship heavily limits the number of possible short female relationships. Our dataset has few labels for minor female roles and thus short, explicitly female-female relationships are hard to find. In addition, though, analysis of the lexical features that predict this suggest that the difference is fairly subtle, more so than a gender divide might suggest: the significant positive indicators of a long relationship with respect to randomly significant are "it," "we," and "we ll", while the negative indicators are "name," "he," and "mr," which suggest that the identification of a collective "we" might show a longer connection but very little else that obviously signals a relationship's length.

5 Related Work

There exists prior work analyzing the differences in language between male and female writing, by Argamon, Koppel, Fine, and Shimoni (Argamon et al., 2003). Herring and Paolillo at Indiana University have shown relations in the style and content of weblogs to the gender of the writer (Herring and Paolillo, 2006). The investigative strategy we use for comparing n-gram probabilities stems from work done by Monroe, Colaresi, and Quinn on distinguishing the contentful differences in language of conservatives and liberals on political subjects (Monroe et al., 2008). Recently, researchers used a simpler version of n-gram analysis to distinguish funded from not-funded Kickstarter campaigns based on linguistic cues (Mitra and Gilbert, 2014).

6 Conclusion

Finding words that are stereotypically male or female came can be done rather quickly and roughly. Yet more sophisticated techniques provide more reliable and believable data. Isolating the right subset of the data to use with proper control methods, and then extracting useful information from this subset results in interesting and significant results. In our small dataset, we find that simple lexical features were by far the most useful for prediction, and that

sentiment and structure prove less effective in the setting of our movie scripts corpus. We also isolate several simpler discourse features that suggest interesting differences between single-gender and two-gender conversations and gendered speech.

7 Acknowledgements

We thank C. Danescu-Niculescu-Mizil, L. Lee, D. Mimno, J. Hessel, and the members of the NLP and Social Interaction course at Cornell for their support and ideas in developing this paper. We thank the workshop chairs and our anonymous reviewers for their thoughtful comments and suggestions.

References

[Argamon et al.2003] Shlomo Argamon, Moshe Koppel, Jonathan Fine, and Anat Rachel Shimoni. 2003. Gender, genre, and writing style in formal written texts. *Text & Talk*, 23(3):321–346.

[Bird et al.2009] Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python*. O'Reilly Media. Available at http://www.nltk.org/book/.

[Cressman et al.2009] Dale L Cressman, Mark Callister, Tom Robinson, and Chris Near. 2009. Swearing in the cinema: An analysis of profanity in US teen-oriented movies, 1980–2006. *Journal of Children and Media*, 3(2):117–135.

[Danescu-Niculescu-Mizil et al.2012a] Cristian Danescu-Niculescu-Mizil, Justin Cheng, Jon Kleinberg, and Lillian Lee. 2012a. You had me at hello: How phrasing affects memorability. In *Proceedings of ACL*, pages 892–901.

[Danescu-Niculescu-Mizil et al.2012b] Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg. 2012b. Echoes of power: Language effects and power differences in social interaction. In *Proceedings of the 21st international conference on World Wide Web*, pages 699–708. ACM.

[Herring and Paolillo2006] Susan C Herring and John C Paolillo. 2006. Gender and genre variation in weblogs. *Journal of Sociolinguistics*, 10(4):439–459.

[Holmes2013] Janet Holmes. 2013. *Women, men and politeness*. Routledge.

[Hutto and Gilbert2014] CJ Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International AAAI Conference on Weblogs and Social Media*.

[Jefferson1988] Gail Jefferson. 1988. On the sequential organization of troubles-talk in ordinary conversation. *Social problems*, 35(4):418–441.

[Lakoff1973] Robin Lakoff. 1973. Language and woman's place. *Language in society*, 2(01):45–79.

[Lauzen2015] Martha M Lauzen. 2015. It's a man's (celluloid) world: On-screen representations of female characters in the top 100 films of 2014. Center for the Study of Women in Television and Film, http://womenintvfilm.sdsu.edu/files/2014_Its_a_Mans_World_Report.pdf.

[Mitra and Gilbert2014] Tanushree Mitra and Eric Gilbert. 2014. The language that gets people to give: Phrases that predict success on kickstarter. In *Proceedings of the 17th ACM conference on computer supported cooperative work & social computing*, pages 49–61. ACM.

[Monroe et al.2008] Burt L Monroe, Michael P Colaresi, and Kevin M Quinn. 2008. Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372–403.

[Pedregosa et al.2011] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

[Pennebaker et al.2003] James W Pennebaker, Matthias R Mehl, and Kate G Niederhoffer. 2003. Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology*, 54(1):547–577.

[Tannen1991] Deborah Tannen. 1991. *You just don't understand: Women and men in conversation*. Virago London.

[Toutanova et al.2003] Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the ACL on Human Language Technology-Volume 1*, pages 173–180. ACL.

[Warriner et al.2013] Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior research methods*, 45(4):1191–1207.