Extracting and Correlating Respect from California Parole Hearing Transcripts with Parole Outcomes

Ramin Ahmari
Stanford University
M.S., B.S. Computer Science
rahmari@stanford.edu

Chris Lucas
Stanford University
M.S., B.S. Computer Science
cflucas@stanford.edu

Michael Smith
Stanford University
B.S. Symbolic Systems
msmith11@stanford.edu

Abstract

We applied techniques from Voigt et al's [1] *Language from Police Body Camera Footage Shows Racial Disparities in Office Respect* in the different domain of parole hearing transcripts. Using their methodologies, we extracted elements of *respectful language* between inmates and the commissioners conducting the hearing. After identifying these features, we correlated them with the parole outcomes in two different experiments. The first experiment was a binary classification task of predicting whether an inmate's request for parole would be granted or denied. The second task aimed to predict the number of years an inmate's request would be denied for where zero would signify the parole request was granted. In the first task, we demonstrate a SVC model that predicts with 87% accuracy the outcome of the trial; in the second task, we demonstrate a linear regression model that predicts the length of parole denial within two years on average.

1 Overview

Our team was inspired by the idea of exploring the criminal justice system and detecting racial biases intertwined in the parole granting system. Identifying the existence of these prejudices is not only something that we as a team are extremely passionate about, but it is also of utmost importance from a social justice standpoint. The rates of incarceration of black and brown communities are a plague to equality and one of the gravest instances of race-based violence.

We were drawn to the work being done by Kristin Bell from Yale University alongside a team of Stanford researchers Catalin Voss and Jenny Hong looking at linguistic features in parole hearing transcripts. We teamed up with them to build off of their work and apply new techniques to identify elements of 'respect'.

2 Problem Definition and Significance

Adjacent to The Innocence Project (https://www.innocenceproject.org/), our project aims to extract higher-level linguistic information in order to identify the sentiment of *respect*. Our data set consists of five-hundred text documents of parole hearing records from the state of California. We believe detecting elements of respect between each commissioner and inmate in these parole hearing documents will be useful features in tandem with features such as race, number of inmate violations, etc. when analyzing the relationship among these variables and the final outcome of the parole hearing. In line with the mission of The Innocence Project, an ultimate goal of this project is to be able to identify and bring justice to wrongly denied inmates.

3 Previous Work

We are using Voigt et. al's *Language from Police Body Camera Footage Shows Racial Disparities in Office Respect* [1] as a cornerstone for our project's architecture and methodology. This paper investigates racial disparities in language extracted from police body camera footage. It finds that police officers speak significantly less respectfully to black than to white community members in everyday traffic stops. Respect is established via computational linguistic techniques that are further explained below.

In their paper, Voigt et al. use police body camera footage from the Oakland police department, particularly from vehicle stops of white (299 stops) and black (682 stops) community members, during the month of April 2014 (981 stops in total).

The footage was then transcribed to provide textual linguistic data. 245 different police officers conducted the stops and the resulting footage consisted of 183 hours, yielding 36,738 usable utterances. The paper split its investigation into three separate studies. In its first study, Voigt et al. investigate perceptions of office treatment from language by testing whether humans can reliably judge respect from officer's language. It was further investigated whether a racial bias was observable in these findings. To establish the complex of respect, utterances were rated along five Likert scales of overlapping notions that contribute to respect - in particular respectfulness, politeness, friendliness, formality and impartiality. Each utterance was rated by at least 10 participants. The reliability of the established perceptions was confirmed via Cronbach αs of 0.73 to 0.91 which indicate moderate to high agreement. A linear mixed-effects regression model combined these different scales to estimate the effect of race across interactions and revealed that utterances from the officer directed toward black members were rates as less respectful, less polite, less friendly, less formal and less impartial compared to utterances from officers towards white members, regardless of gender and age.

Voigt et al. [1] then proceeded to apply Principal Component Analysis (PCA) and found that two principal components explained 93.2% of the data variance. The first component explained 71.3% and has positive loadings mainly on the impartiality, respectfulness, friendliness, and politeness dimensions. This component was characterized as the broader concept of respect while the second components, attributing to 21.9% of data, had very strong positive loadings on the formality dimensions and weak negative loadings on the friendliness dimension and was characterized as the concept of the Formality. Therefore, formality and respect were established as distinct concepts.Applied back to mixed-effects model as outcome variables, Voigt et al. find that officers employ equal formality across drivers but higher respect in white drivers.

In the second part of their study, Voigt et al. investigate linguistic correlates of respect in order to scale their findings and provide a general, computational solution to the analysis of body camera data. Voigt et al. thus develop computational, linguistic models of respect and of formality which are based on linguistic theory. Accordingly, respectful language thus includes language such as apologies and agency-giving. Voigt et al. reference the papers by Parabhakaran, Rambow and Diab (2012)[2], Tausczik and Pennebaker (2010)[3], Danescu-Niculescu-Mizil, Lee, Pang and Kleinberg (2012) [4] and Danescu-Niculescu-Mizil, Sudhof, Jurafsky, Leskovec, and Potts (2013) [5] for their feature extraction regarding this dataset. Their counts are then log-transformed and fed into linear regression models predicting the perceptual ratings of respect and Formality as previously established, assigned ratings that largely agree with the human-obtained ratings (root-mean square of 0.840 compared to 0.842 for Respect and 0.882 compared to 0.764 for Formality). These results thus prove that formality and respect can be deduced from linguistic features.

The third part of the paper then investigates these results in terms of racial disparity. When split by race outcomes, informal titles, the demand to put the hands on the wheel, asks for agency and linguistic negation were much more common in stop of black people compared to a higher amount of talks about concern for citizen safety, giving agency, formal titles and reassurance in stops of white people. Particularly the concern for safety, reassurance and agency indicate a lack or presence of respect.

These models are then applied to the entire corpus to generate predicted scores for a linear mixed-effect model for formality and respect. After accounting for covariates such as race, age, gender, search conductance and result of the arrest, utterances spoken by officers to white people (and older people) score higher in respect. Officer race was not significant. Furthermore, offense severity was not predictive of officer respect levels. Race (in contrast to gender and age) was also not associated with the formality of officers' utterances. There was a normal distribution of respect among all officers and thus the difference in respect cannot be explained by a small number of officers. Officer respect also decreased throughout the interaction with black drivers while it increased with white drivers. White community members are 57% more likely to hear officers say one of the most respectful utterances while black community members are 61% more likely to hear one of the least respectful utterances. More importantly, racial disparities can be observed in the first 5% of

a conversation, indicating preconceived notions at play.

Voigt et al.'s paper [1] clearly shows us that respect can be deduced clearly from linguistic features and, given the preceding papers efforts, we aim to replicate and expand upon these methodologies to establish a framework for measuring respect within parole board hearings to investigate its racial disparities and effect on outcomes.

4 Methodology

Our architecture for identifying levels of *respect* in parole hearing interactions between commissioner and inmate is based on a series of underlying features that we must first identify. Similar to Voigt et al's approach with body camera footage and its transcription, we want to parse PDF transcript documents and extract semblances of *politeness*, *friendliness*, *formality* and *impartiality* as well as the counts of several LIWC-related features. However, given our project resides in a slightly different domain, we must adapt Voigt et al's methodologies as well as the methodologies for the four underlying papers as the foundation for our project (as mentioned in depth in our literature review).

From our team of researchers Catalin Voss and Jenny Hong, we have preprocessing and parsing pipelines to manage digitizing the parole hearing transcripts into documents which we can then break down into a collection of statements. For example:

```
docs = session.query(ReconDocument)
sections = docs[0].sections
statements = sections[0].statements
sentences = statements[0].sentences
```

4.1 Data

Originally, our dataset consisted of 541 parole hearing transcripts after we filtered out approximately thirty documents that were either repeats of hearings already in the dataset or corrupted. Each document averaged over one hundred pages; our shortest was approximately fifty pages, our largest approximately three hundred. Collectively, the 541 documents account for over 1.6 million sentences/sentence fragments that we were able to use in our feature pipelines. However, after we wrote a script to extract the parole decisions (GRANTED/DENIED), we found that 70 documents either did not have a decision (for a variety of reasons - illness, deportation, etc.) or did not have a decision that was explicitly stated in the text; we removed those 70, and were left with 471 documents that contain nearly 1.4 million sentences/sentence fragments. On each document, we extract thirty-two features: twenty based on the LIWC dataset (ten for both inmate and presiding commissioner), and the remaining twelve (eight for the presiding commissioner, six for the inmate) based on features of politeness and power found in Voigt et al. We expand upon why we chose these features in subsequent sections.

In the spirit of incremental development and testing, we first tested our model on a subset of approximately fifty-percent of our dataset (261 out of 471 documents). We found that maintaining document-level counts of each feature, and then averaging the counts for each feature across all documents was helpful to normalize differences in document length. After we generated thresholds for each feature, we binarized our dataset such that a feature i in a document would be 1 if its count was greater than or equal to the threshold value, and 0 otherwise. After we determined that a binarized dataset was beneficial, we then expanded to use the full dataset.

Our datasets and code are available upon request.

4.2 Politeness

For scoring the *politeness* of a particular document, we previously mentioned that Voigt et al. used ten Mechanical Turk workers to subjectively rate each document and then averaged those together to produce an intermediary politeness rating. They also layered on a lexicon based approach which consisted of searching each document for certain keywords and phrases to augment this politeness score. Given our limited timeline and monetary budget for our course project, we forewent the MTurk approach and developed a similar lexicon-mapping based algorithm to measure politeness. We rely on the same linguistic lexicons to measure *apology*, *ask-agency* ("can I", "may I", etc.), *give-agency* ("you can", "you may", etc.), *gratitude*, and *permission*. We count the number of total utterances, one count for the commissioner statements and one count for the inmate statements and then compare these counts to an average threshold and feed this information into the same system we previously described in Voigt et al [1].

4.3 Linguistic Inquiry and Word Count (LIWC)

Similarly to Tausczik and Pennebaker (2010) [3], we use Linguistic Inquiry and Word Count (LIWC) to count words in psychologically meaningful categories (such as social relationships and emotionality) that are able to capture interpersonal dynamics. Given that we are of a constrained monetary budget, we are using the 2007 LIWC dictionary provided by Chris Potts in the Piazza forum.

We followed Tausczik and Pennebaker's heuristics for establishing status, dominance and social hierarchy by looking at plural / singular use but have found that the 2015 LIWC dictionary contains a power sub-dictionary that would be much more powerful in our analysis. We are reaching out to our project group to see if we can get the LIWC 2015 library funded for the purposes of our project.

Besides the emotional categories, we furthermore replicated the psycholinguistic reasoning in Leshed, Hancock, Cosley, McLeod, and Gay (2007) [6] and Kaceqicz, Pennebaker, Davis, Jeon, and Graesser (2009) [7] to count the first-person singular as well as overall word count which were postulated to be indicative of a domineering position and social hierarchical imbalance.

Furthermore, we investigated group cohesion by the use of first-person plurals as postulated by Sexton and Helmreich (2000) [8]. By measuring the frequency of tentative and filler words via LIWC, we were also able to establish a measure of prejudice as filler words indicate tentative language and unformed opinions.

4.4 Overt Displays of Power

Given the close relationship of respect with social constructs of domination, hierarchy and power (for example, Voigt et al. has formality - a part of respect - as one of their principal components in their analysis of their data), we thought that a closer look was warranted into linguistic displays of power. We initially thought that Voigt et al. directly drew on Prabhakaran, Rambow and Diab (2012) [2], who used a small corpus of the widely accessible Enron dataset annotated with dialog acts to predict power differentials, so we recreated their SVM model using the Enron emails. Like Prabhakaran, Rambow and Diab, we utilized tokenization, POS tagging, lemmatization as features into a linear (C=1) support vector machine. We were able to approach within .1 of their F1-score. However, in ablative analysis of the features, we found that, when dialog act features are removed, the F1-score drops precipitously by over 20% in absolute terms. Prabhakaran et al. (2012) [2], in their analysis, found similar results even when they used an automatic dialog act tagger from Hu et al. (2009). When we realized this, we removed overt displays of powers from consideration as one of our features.

We later learned that Voigt et al.[1] did not use ODPs in their work, but Voigt and our team thought that it would be of great interest to utilize this line of methodology; we foresee that determining power differences or perhaps abuses of power within the parole board hearings on the side of the commissioner would be a valuable feature in its lack of respect. To begin a proxy for this, we decided to maintain a count of the number of times the presiding commissioner said the first and the last name of the inmate. Voigt et al. found that use of the forename was correlated with a lack of respect, and use of the surname was correlated with more respectful language. Our findings thus far suggest that neither feature is as seminal in our domain (the two features exist among the bottom fifteen features in our importance hierarchy) as it is in the traffic stop setting. This follows, we believe, from the more formalized and structured setting of the courtroom.

4.5 Feature Unification

This next step involves synthesizing everything we extracted from our three pipelines into one predictive model. We concatenated everything gleaned from each hearing into a single, document-level feature representation. In total, this gave us 32 features – twenty from the LIWC pipeline, ten from the politeness pipeline, and two from the power pipeline.

5 Preliminary Results

By first developing on a smaller subset of the data, we were able to gain some quick results and, more importantly, some important insights before we transitioned to use the entire collection of documents.

5.1 Experiment on Small Data Subset

We ran our model on a subset of 261 documents to get an initial sense of performance.
We investigated both regular classification metrics, such as the F1-score of our models, as well as the feature importances in predicting parole outcomes as a binary classification problem – granted versus denied.
For this subset of our data, the top three features were the commissioner's negativity, inmate's selfishness, and the commissioner's confrontational attitude. Each of these features came from the LIWC feature pipeline that we had established.
Additionally, our classification models were compared against a two baselines we established: a random classifier yielding 50% accuracy, and a slightly more advanced baseline that always predicted denied, yielding a base rate of 80%. We felt this latter baseline would be a stronger, more difficult benchmark to use given the disproportionate nature of our dataset being heavily skewing towards parole denial.
We achieved a 100% accuracy with our Decision Tree Classifiers and generally found classification curacies of around 90% for tree-based classifiers.
Initially elated given that a high number could indicate a very strong performance, we were nonetheless surprised by the strength of the generalizability this model already had at this stage and investigated further.
We also saw understandably high variance in our feature importance results. This is something we hoped to solve once our model ingested more data. However, we were also able to see that LIWC features consistently ranked higher in importance compared the features derived from the politeness pipeline.

5.2 Extensions and Insights from Further Investigation

Investigating the surprising effectiveness of the 95+% accurate Decision Tree Classifiers, we statistically investigated the simple count characteristics of our dataset and realized that the dataset was heavily skewed in terms of binarized outcome, with a heavy skew to 0 or denial of parole towards the beginning of the dataset.
We therefore decided to run 10-fold cross-validation with a 5-split moving forward to account for this possibility, In fact, when running our previous model, our model accuracy dipped into the 70%s.
One enhancement we made to our model consisted of modifying some of the phrases used in our lexicons to better align with the verbiage used in parole hearing transcripts. Note that the lexicons we were leveraging up until this point were derived from Voigt et. al's work where the dataset was transcribed dialogue between officer and civilian from camera footage. Given the nature of our dataset, we read through a small subset of the incorrectly classified hearings to identify more common phrases to use.
The main takeaway from this exercise was the difference between colloquial and formal language. In the case of the camera footage, officer-civilian relations were more casual in tone and phrasing compared to that of an official court proceeding. Phrases such as 'oops', 'my bad', etc. to signal signs of apology were not as relevant as more formal language such as "i apologize', 'my apologies', and 'forgive me.'
Additionally, we extracted the actual number years of until the next hearing when parole was granted which would give us a more nuanced predictive metric. The problem statement therefore lend itself naturally to the application of regressions models.

5.3 Experiment on Large Dataset

Having sample-run our models on a smaller dataset and having investigated the results afterwards to refine our model as described above, we moved onto the larger dataset of 471 documents that is currently available to us.
Figure 1 and figure 2 show our results from our classification task.

Note that for the larger dataset, our denied classifier baseline rate is 0.68. The feature rankings are shown in Figure 2 – note that _C denotes a commissioner feature and _I denotes an inmate feature.
As evident in Figure 1, simpler models such as the SVC and Logistic Regression greatly outperformed both of our baselines and additionally strongly outperformed the decision trees as well. Our resultant model accuracy was strong at 87%. It, furthermore, confirmed our intuition regarding the misleading perfect or near-perfect classification accuracy we had established with on the smaller data subset with a static partition of the

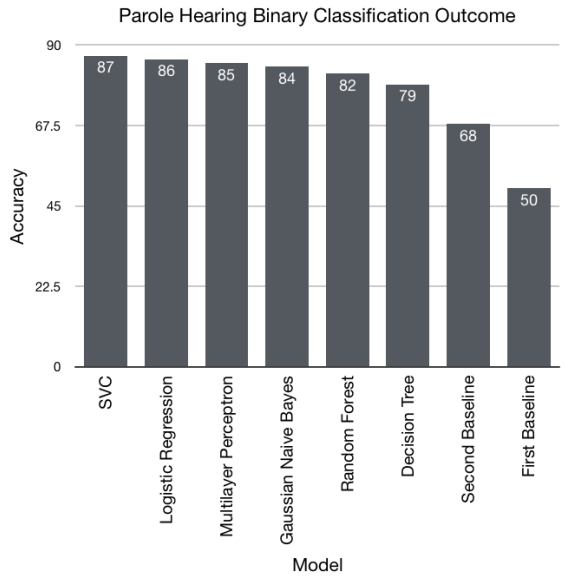


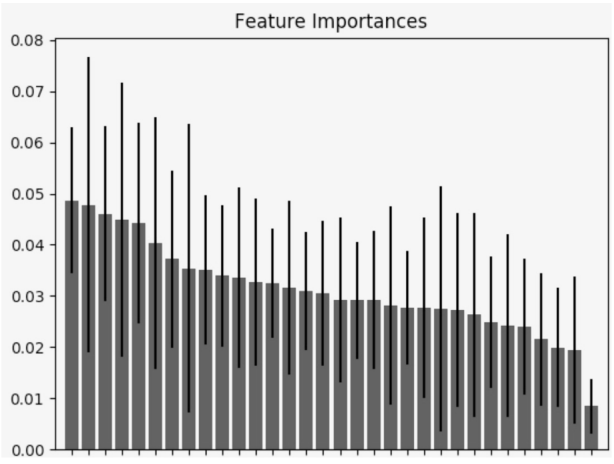
Figure 1: Large Dataset Model Accuracies

dataset that played into its inherent skewedness.

The variance of our feature importances, while still quite variable, stabilized after our model was run on the larger dataset. From Figure 2 we can see that the inmates' positivity as well as the commissioners' selfishness and tendency to give agency were the three most important features.These are quite different from the volatile feature variances identified in the smaller subset and showcase how variable our first estimations were.

Linear Regression proved most fruitful in the prediction of the number of years until the next hearing for an inmate whose parole was denied. This was a more nuanced classification task of our previously binarized – and thus simplified – problem statement. As evident in Figure 3, at the mean squared error was around 4.6. This estimate of a 2 year mean inaccuracy in prediction of the years of until a next parole hearing was strong given that the range of years until a next parole hearing was up to 15 years.

Similarly to the classification results, we were able to investigate the importance of our features and realized that once again these importances were quite variable and only provided a weak confidence. The identified top three most important features were the tendency of the commissioner to give agency, their friendliness and the amount of apologizing nature in the inmates' speech. While feature importances were quite variable, interestingly the top feature, the tendency of giving agency by the commissioners, usually remained within the top three and can also be found within the top three of the classification model.



Positive_I	0.048678	Perceptive_C	0.030570
Selfish_C	0.047826	Positive_C	0.029257
GiveAgency_C	0.046004	Negative_I	0.029154
Social_I	0.044961	RudeAggressive_C	0.029141
Friendly_C	0.044234	Negative_C	0.028084
AskAgency_I	0.040324	Perceptive_I	0.027734
Indecisive_I	0.037182	Apologize_I	0.027640
Gratitutde_I	0.035403	Selfish_I	0.027461
Please_I	0.035131	Please_C	0.027283
Social_C	0.034005	GiveAgency_I	0.026294
AskAgency_C	0.033574	Confrontational_C	0.024916
Emotive_I	0.032708	Rude/Aggressive_I	0.024205
Indecisive_I	0.032497	Apologize_C	0.021516
Emotive_C	0.031529	Confrontational_I	0.019918
Friendly_I	0.030977	Gratitude_C	0.019329

Figure 2: Large Dataset Feature Importance

5.4 Conclusion

We are very pleased with our results and we think that using these features for more complex analysis on the parole hearing transcripts could prove fruitful.

For example, a sub-team within the team of researchers at Stanford developed a pipeline to learn an inmate's *parole hearing status* (is this an initial hearing or subsequent hearing) and an inmate's *future employment status*. Synthesizing the work from their team with our own work could be an interesting next step. The first possibility to consider would be concatenating the extracted features whereas another, potentially richer possibility would be to examine the relationship between the two projects. We could look at correlations

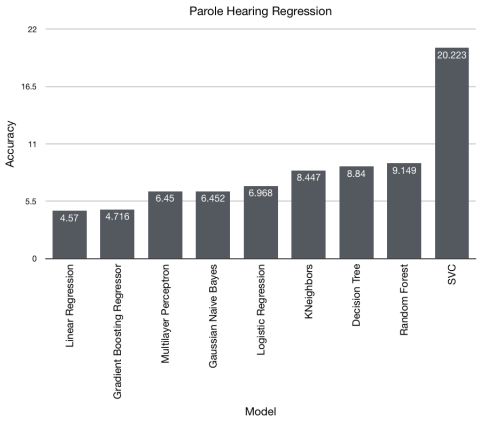
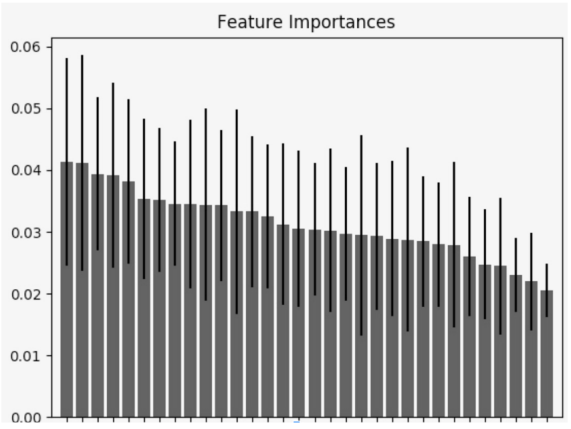


Figure 3: Large Dataset RMSE

among our features and their features, essentially asking questions like is an inmate's future employment status a predictor of how respectful a commissioner is in a particular hearing.

By late June, we expect to gain access to an expanded dataset of 50,000 parole hearings as soon as Kristin Bell at Yale receives them from the State of California. On this corpus of likely over 150 million sentences, we anticipate that we will continue to improve our models and be able to understand trends in demonstrated respect. We are particularly excited to address the remaining issues of high variability in the identification of feature importances with the addition of data. When we receive race data on the inmates, we plan to analyze the outcomes to see if there are any differences in treatment like those found in Voigt et al. [1]



GiveAgency_C	0.041358	GiveAgency_I	0.030464
Friendly_C	0.041146	Confrontational_I	0.030386
Please_I	0.039393	Confrontational_C	0.030242
AskAgency_I	0.039232	Selfish_I	0.029615
AskAgency_C	0.038134	Rude/Aggressive_C	0.029429
Emotive_I	0.035324	Please_C	0.029292
Positive_I	0.035225	Negative_I	0.028877
Social_C	0.034562	Social_I	0.028739
Perceptive_I	0.034512	Selfish_C	0.028433
Apologize_I	0.034412	Apologize_C	0.027976
Indecisive_C	0.034322	Rude/Aggressive_I	0.027902
Positive_C	0.033290	Negative_C	0.026002
Perceptive_C	0.033277	Gratitude_I	0.024429
Indecisive_I	0.032506	Friendly_I	0.023081
Emotive_C	0.031235	Gratitude_C	0.021978

Figure 4: Large Dataset RMSE

References

[1] Voigt et al. 2017. *Language from police body camera footage shows racial disparities in officer respect*, Proceedings of the National Academy of Sciences

[2] Prabhakaran V, Rambow O, Diab M 2012. *Predicting overt display of power in written dialogs.*, Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics Assoc Comput Linguist, Stroudsburg, PA

[3] Tausczik and Pennebaker 2010. *The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods*, Journal of Language and Social Psychology 29(1) 2454

[4] Danescu-Niculescu-Mizil, et al. 2012. *Echoes of Power: Language Effects and Power Differences in Social Interaction*, Proceedings of the 21st International Conference on World Wide Web Assoc Comput Mach, New York

[5] Danescu-Niculescu-Mizil, et al. 2013. *A computational approach to politeness with application to social factors.*, Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics Assoc Comput Linguist, Stroudsburg, PA

[6] Leshed, Hancock, Cosley, McLeod, and Gay 2010. *Visualizing Language Use in Team Conversations: Designing through Theory, Experiments, and Iterations*, CHI 2010

[7] Kacewicz, Pennebaker, Davis, Jeon, and Graesser 2009. *Pronoun Use Reflects Standings in Social Hierarchies*, Journal of Language and Social Psychology

[8] Sexton, Bryan J., Thomas, Eric J., Helmreich, Robert L. 2000. *Error, stress, and teamwork in medicine and aviation: cross sectional surveys*, BMI