Gendered pronoun usage in a Corporate Environment

Zubair Abid, 20171076

Abstract:

We investigated the usage of gendered vs neutral singular pronouns, and its distribution across gender and power lines, in a corporate setting using the Enron Email Corpus. It uses a freely available extension to the Enron Email Corpus where genders of senders of 87% messages are reliably identified. Using this data, we test a hypothesis drawn from sociolinguistic literature pertaining to gender and power: People in positions of authority use more gender-neutral pronouns (or gender non-specific pronouns) in general, and women across power lines do so in particular.

Introduction:

Contemporary English comprises two classes of third person singular pronouns divided by gender, that is male and female (he/she, his/hers, etc). In order to refer to a person of unknown gender, people typically use combinations of both forms, like *he/she*, *(s)he*, and so on, or use the generic male form (*he* and variations). In recent times the usage of the third person plural *they* as a singular pronoun to refer to a third person of unknown gender has increased, for a multitude of reasons - to avoid the bulkiness associated with compound pronouns like *he/she*, to be more inclusive of non-binary genders, and because of the fact that such usage is not particularly alien to standard English.

It has been observed that men and women communicate differently in different contexts. (Prabhakaran et al., 2014; Herring, 2008) showed that female superiors tended to use "face-saving" strategies at work, including conventionally polite requests and impersonalised directives - in comparison to male superiors. It was also seen that there was little to no difference in the language use between female superiors and subordinates, but male superiors showed more Overt Displays of Power as compared to male subordinates. Based on these results we aim to see if power hierarchy and gender differences in a corporate environment like the Enron Corporation make any differences when referring to a generic third person.

The paper is structured as follows. Section 2 deals with related work, including some that inspired this paper and some of the experimental methods used in them that we could adopt and improve on. The Hypothesis is formalized and tested in Section 3, with the experimental method described in detail. We then turn to the analysis of the data in Section 4. We then conclude and discuss possible future work/improvements.

Related Work:

There is a lot of sociolinguistic work done on gender and language use, some of it specifically related to language use in a work environment (Kendall and Tannen, 1997; Holmes and Stubbe, 2003; Kendall, 2003; Herring, 2008, Prabhakaran et al., 2014). We won't discuss each of these at length because of space constraints, but will pick on and highlight the relevant/influential ones here.

One paper that has to been particularly influential to the conceptualization of this project is Holmes and Stubbe (2003). In it, through case studies in the country of New Zealand, they establish two conclusions - first, that female managers tend to break many of the stereotypes of 'feminine' communicative methods. Second, that while as equally effective as their male counterparts, they use different strategies to keep control of the discourse ("they also use a wide variety of more subtle strategies to keep control of the discourse, with choice of strategy influenced by specific context."). This includes more directness in meetings while subordinating tasks, but a less confrontational style when dealing with problems on a one-to-one basis. A bulk of our hypothesis, that women would use more gender neutral pronouns to refer to an unspecified singular person in order to appear less direct and thus more accomodating, is based on the second conclusion.

Another rather influential paper was Prabhakaran et al. (2014) - the source of the Gender Identified Enron Corpus (GIEC) used for the experiment. In addition to that, the results of this paper corresponding to conclusions in Herring (2008), provided similar impetus to the hypothesis as the Holmes and Stubbe (2003) results. One thing to note is that unlike in either of these papers, the concept of a Gender Environment has not been used in this experiment/analysis.

Hypothesis and Experiment:

In this subsection we discuss the hypothesis and the experimental method used to check for its correctness. Data for the experiment was obtained from the GIEC, a corpus of approximately 280,000 emails of employees of the now defunct Enron Corp. It is tagged with each sender's gender and position in the organizational hierarchy.

Since 280,000 emails are a bit hard to read through all at once, we had to run the data through some filters based on the hypothesis constraints.

- **Hypothesis:** *People in positions of authority use more gender-neutral pronouns (or gender non-specific pronouns) in general, and women across power lines do so in particular.*

We also automated as much of the pipeline as possible, which includes filtering the data through a pronoun matching wordlist, running a heuristic to increase frequency of relevant data match, and running a pattern matcher to simplify match searches for final annotation. The final annotation had to be done manually.

The pipeline comprised of four basic sections - a filter stage to remove any data obviously irrelevant to the analysis/hypothesis, a heuristic filter to make manual matching easier (speeds up manual annotation by a factor of ~30), and a final annotator helper that highlights the relevant heuristic word and displays possible pronouns for possible coreference resolution. The fourth and final stage (after manual annotation) takes all the final data selected and runs it through the original database again to get gender and hierarchy data of the sender of the email. This is run after the rest of the pipeline as multiple expensive database calls are made, increasing the runtime of the algorithm.

The pipeline can be seen in https://github.com/zubairabid/LingoProject/tree/master/pipeline

The filter: Since we're running the experiment to check for instances of third person singular pronouns, we run a filter to eliminate all emails not containing any such relevant pronouns - these emails are treated as noise and not considered for further analysis (a stub explaining the shortlist criteria and some relevant code is given below). This also translates the bulk of the data required from MongoDB JSON objects to simple text files.

Relevant pronoun list = ['he', 'she', 'they', 'her', 'hers', 'his', 'him', 'theirs', 'them', 'their'].
Code:

```
for key in plist:
	# eliminating non matches
	if (re.search(r'\b' + key + r'\b', origin) is None):
		continue
	ll.append(key)
```

Heuristic filtering: After this, emails are manually marked as relevant or not and annotated based on existence of valid data (i.e, referrals to third persons in the singular). Since this data is sparse and there are upwards of 5000 emails to still read through, after a sampling of 7 annotations a heuristic was developed to increase the hit rate of relevant data. It reduces the number of emails to only about 1300, of which less than 10% contain information needed. This necessitates the exclusion of useful data not considered in the heuristic, but since 47 emails are being picked at random anyway, this shouldn't affect the numbers by much. A description of the heuristic used is given below.

Heuristic: At least one of ['person', 'customer', 'witness', 'manager', 'each'] must be contained in the emails being considered.

Annotator helper: a basic python program to highlight all instances of a heuristic in an email body. Helps speed up manual annotation of the data.

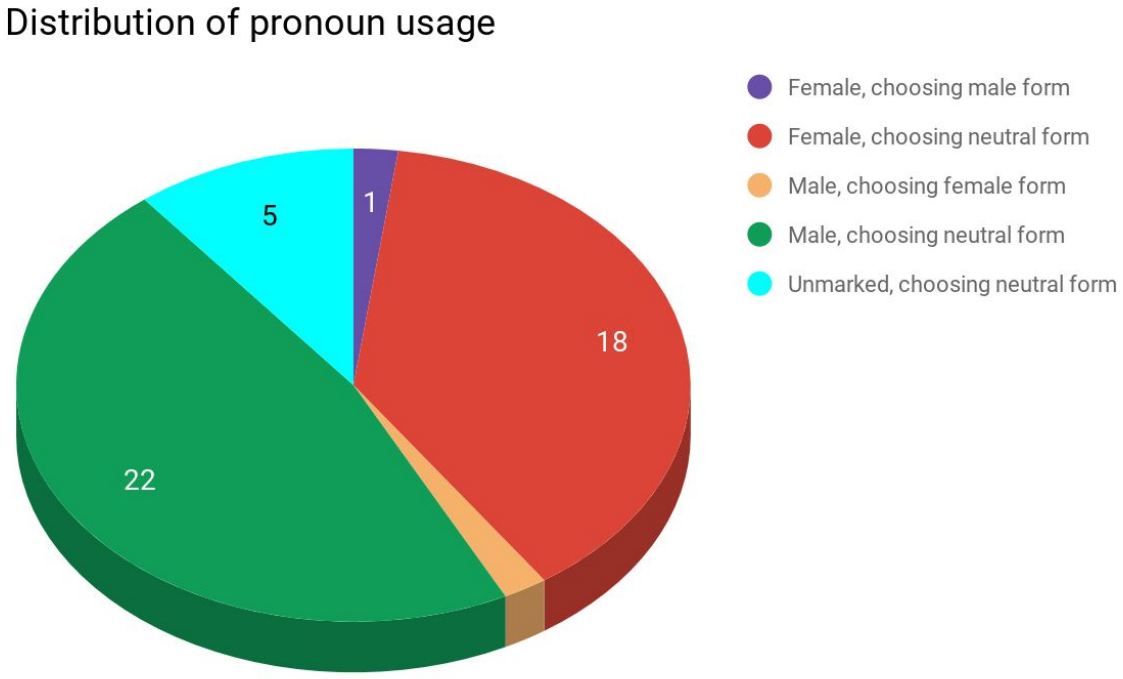
Sender Information Extractor: Is run after the manual annotation portion and the rest of the pipeline is complete. It makes a query to the entities table in the database based on sender uid extracted from the emails table and gets gender and affiliation results. If either doesn't exist, it's marked as 'UNMARKED DATA'

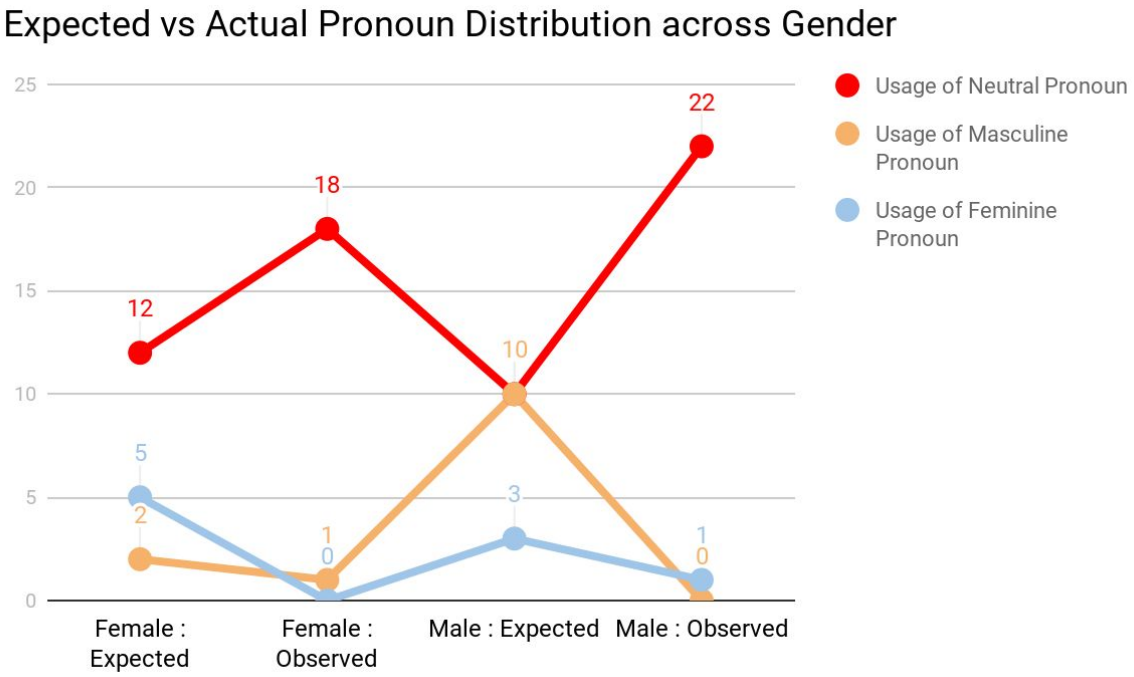
Altogether, 47 emails were annotated for further analysis.

Results:

Usage of neutral, masculine, and feminine pronouns across given social categories

Gender →	**Female**				**Male**				**Unknown**
Affiliation →	**Core**	**Non Core**	**Non Enron**	**Unmarked**	**Core**	**Non Core**	**Non Enron**	**Unmarked**	**Unmarked**
Using neutral pronoun		3	2	13		4	3	15	5
Using masculine pronoun				1					
Using feminine pronoun						1			





Conclusion:

Our hypothesis is thus invalidated; it seems that barring a few outliers, every reference to third person singulars in the Enron Email Corpus is gender-neutral.

Some reasons for this result are obvious - there are often repeated instances of the noun being used over and over again instead of reverting to the pronominal form (particularly noticeable in the word ‘customers’). It is (seems to be) an outcome of formality in text, something that would probably not happen in regular speech.

This highlights another possible reason for the lack of gendered terms - email, especially in and up to 2004, is a very formal medium of communication, even more so because it’s corporate email (related: Peterson et al., 2011). We hypothesize that data from news forums or chat rooms from around the time would have had vastly different results, which is a possible future study option to examine. For such a study gender data would have to be majorly ignored because of lack of gender data to work with.

Because of poorly defined categories and monolithic distribution, it seems as if hierarchy is not a factor here. Either.

Acknowledgements:

I’d like to thank a bunch of people in this. Professor Dipti M. Sharma, for suggesting the topic. Anirudh Dahiya, for suggesting I could use a heuristic to reduce manual annotation. Deepti, for helping annotate the data. Sayar, for general suggestions throughout. Shelly, for not killing me while I whined about the project.

References:

Janet Holmes and Maria Stubbe. 2003. feminine workplaces: stereotype and reality. The handbook of language and gender, pages 572–599.

Shari Kendall and Deborah Tannen. 1997. Gender and language in the workplace. In Gender and Discourse, pages 81–105. Sage, London.

Shari Kendall. 2003. Creating gendered demeanors of authority at work and at home. The handbook of language and gender, page 600.

Kelly Peterson, Matt Hohensee, and Fei Xia. 2011. Email formality in the workplace: A case study on the enron corpus. In Proceedings of the Workshop on Language in Social Media (LSM 2011), pages 86–95, Portland, Oregon, June. Association for Computational Linguistics.

Prabhakaran, V., Reid, E.E., Rambow, O.: Gender and power: How gender and gender environment affect manifestations of power. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1965–1976. ACL (2014).

Susan C Herring. 2008. Gender and power in online communication. The handbook of language and gender, page 202.