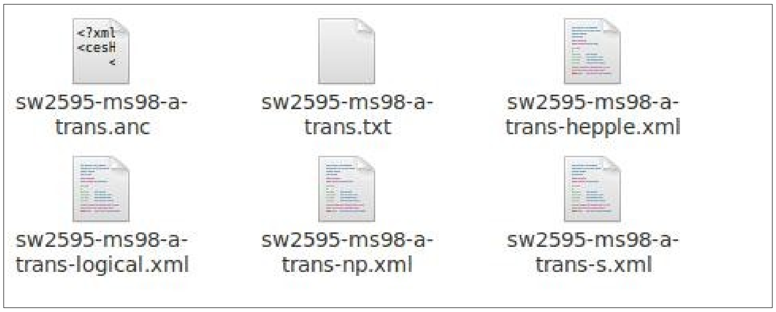
Jeff Killian
December 22nd, 2010
Computational Linguistics
Final Project

One Hell-of-a-Project: Analyzing Swear Frequency by Gender

It is impossible to be a college student and not hear swears; a claim that holds especially true as finals time approaches. But how much do we actually know about swears? In attempting to decide on a topic for my final project, I overheard a discussion regarding swears between one of my roommates and someone of the opposite gender. This sparked my interest in the notion that I might be able to test the claim that, 'one gender swears more frequently than another.' Furthermore, if this were true, is there a difference in the relative frequency of swears? Do females say 'bitch' more often than do males? Not only was this a possible final project, but it was a project that was directly relevant to my life, for swears are plentiful in a college atmosphere. I set out in determining, given that a swear was uttered, if we can make any assumptions as to the gender of the speaker. I wanted to know when cursing, which gender uses which words. My initial hypothesis was that men would swear much more than women across all swear words.

METHODOLOGY

For this project, it was crucial that I pick a fitting corpus. One crucial aspect was that the corpus had to have a verbal section, for I only wanted to know the frequency of swearing in conversation. The main criteria that I desired was that it had enough swears, and had the relative files documenting who stated the swears. After viewing various corpora, I decided that the American National Corpus was a fitting corpus to use. Furthermore, because I was only focusing on swears that were verbally stated, it was fitting that I use a corpus which has a specific 'spoken' section. I therefore narrowed my data down to all the conversations contained within the 'spoken' section of this corpus. Ultimately, I ended up going through approximately 2,400 conversations. For each conversation within the corpus, there were six corresponding files.



sw2595-ms98-a-trans.anc
sw2595-ms98-a-trans.txt
sw2595-ms98-a-trans-hepple.xml
sw2595-ms98-a-trans-logical.xml
sw2595-ms98-a-trans-np.xml
sw2595-ms98-a-trans-s.xml

Illustration 1: Example files corresponding to a single conversation in the corpus

Of these six files, I was only concerned with three: *.anc, *.txt, and *-logical.xml. Within the *.anc file, there was information on the gender and age of each of the participants in the conversation. Furthermore, the file assigns a unique id to each participant-something that turns out to be useful in determining who was speaking at what time. The *.txt file has the actual raw transcript of the conversation. The *-logical.xml file was the most important file. It has information on every utterance and turn-who speaks and when they stop speaking-within the corresponding conversation. It was in this file that I started my searching. For every utterance mentioned in this file, I would go through the conversation file (*.txt) and extract that utterance, searching for a swear within it. If a swear was found, the program then determines who was speaking at that time, and records the various important information about the speaker that swore. All of this information is then pooled, and is returned at the end of the program in a table which shows the counts of swears by gender.

A crucial step of this project is determining what truly counts a swear. This project searched for the following generally accepted swears: fuck, bitch, bastard, hell, shit, crap, damn, and ass. Through a carefully constructed regular expression (included below), the program also accounted for various inflections of these forms.

```
'(?:^|[^a-z])((?:fuck|bitch|bastard|hell|shit|crap|damn|ass)(?:[pt]?
ed|e?s|[tp]?er|[tp]?ing)?) (?=[^a-z])'
```

This string will match the words 'fucks', 'shitting', 'crapper', and 'bitches.' When the program finds something that qualifies as a swear, it records that it found the word and determines who spoke it and their gender and age. All of this information is stored within dictionaries, and is then presented at the end of the program.

Initially, I was only looking at overall swear frequency. However, it occurred to me throughout the project that (1): the number of males and females who speak might not be relatively even, and (2):males and females might not speak the same amount of words. However, my initial proposition- wondering whether males or females said certain swears – could be strongly skewed by this data. For instance, I realized that if males tended to speak more than females, even if we assume that the number of males who spoke and females who spoke was equal, then males, overall should have have higher swear counts, which would skew our data. Similarly, if we hold constant the idea that men and women speak the same amount of words but we have many more men speaking than we do women, then the data will again be strongly biased towards males swearing more, when in reality, any given male could swear the same amount as a women. To account for this, in my final project, I took the question one step further, and calculated the overall number of words spoken by men and spoken by women. Using the relationship between swears spoken and non-swears spoken, I was able to obtain a chi-square value that would be statistically interpretable.

OUTPUT

The program outputs a list of the following for each swear it encounters:

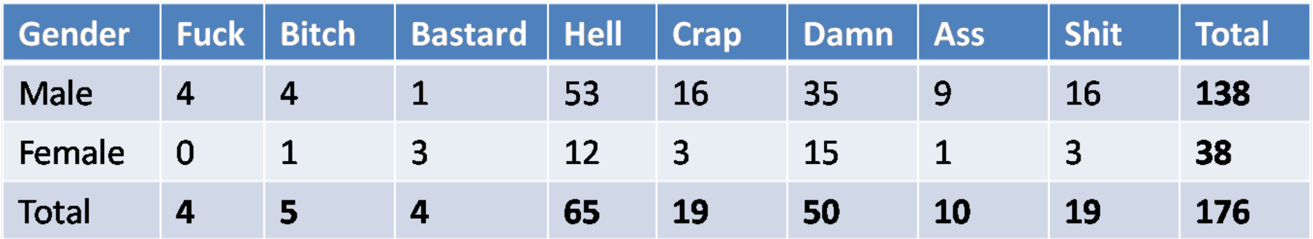
ID Swear {Information on the person that swore} Age
A snippit of the output is included below:

```
B shit {'age': '1936', 'id': 'spkr1239', 'sex': 'M'} 61
B damn {'age': '1944', 'id': 'spkr1443', 'sex': 'F'} 53
A hell {'age': '1940', 'id': 'spkr1436', 'sex': 'M'} 57
A hell {'age': '1952', 'id': 'spkr1359', 'sex': 'M'} 45
B hell {'age': '1944', 'id': 'spkr1159', 'sex': 'M'} 53
```

Once it goes through all of the possible files, the program returns a table with the various totals of swear counts by gender, as well as various information on the overall number of files and people examined.

RESULTS

Overall, there were a total of 2,400 conversations examined. Each conversation has two speakers. There were 2,412 and 2,380 male and female speakers, respectively. There were eight speakers whose gender was undetermined. These eight data points were ignored, for they did not skew the data one way or another, and the results were unable to be analyzed by gender. Of the 3,100,208 total words spoken by males and females, males spoke 1,546,550 (49.8%), while females spoke 1,551,069 (50.03%). Those with gender 'undetermined' spoke less than .1% of the words. What this means is that men and women spoke roughly the same amount (both in number and in word count). We would therefore expect that, if the swear frequencies were the same between genders, males would utter the same amount of swears that the females uttered. The results are reported in the following table.



Gender	Fuck	Bitch	Bastard	Hell	Crap	Damn	Ass	Shit	Total
Male	4	4	1	53	16	35	9	16	**138**
Female	0	1	3	12	3	15	1	3	**38**
Total	**4**	**5**	**4**	**65**	**19**	**50**	**10**	**19**	**176**

Illustration 2: Swear Frequency by Gender

There were both positive and negative aspects of my results. First and foremost, they seemed to support my hypothesis that men tended to swear more than women across all swear words, the only exception of this being the word 'bastard', in which females said three times to the males' one. Looking at the raw count of swears, we see males swore 138 times, whereas females swore a total of 38 times. Because males and females spoke roughly the same amount of words (1546550 and 1551069, respectively), this tends to show that , given a subset of words spoken by each gender, the male's should be expected to have more swears. One negative aspect of the results obtained was that there were not as many swears within the conversations as I would be preferred. Because of this, it becomes extremely difficult to perform chi-square tests on the individual cells, for most have expected cell counts below 5.

Looking at the table in the general, we see a general pattern that males tend to utter more swears verbally than females. However, statistical tests are much more telling; through these we are able to determine if there is a statistically significant χ^2 value, and thus if there is a true difference between males and females. We are able to run a χ^2 test for independence on several variables. Performing this on the counts of the word 'hell' we received, we obtain a χ^2 value of 25.981, which is much outside the rejection region. It therefore seems that men do say 'hell' more than women. Running a similar χ^2 test on the values of 'damn', we obtain a value of 8.058, which is still outside the rejection region. It thus appears as though men say 'damn' more than women. However, the most interesting results came from analyzing swears overall.

χ^2 Test on Overall Swears	**Swears**	**Non-Swears**	**TOTALS**
Males	138	1546412	**1546550**
Females	38	1551069	**1551107**
TOTALS	**176**	**3097481**	**3097657**

Analyzing the overall swears vs. non-swears by gender, we obtain a significant chi square value of 57.116. Therefore, it appears as though the frequency of swearing between males and females is not the same. Unfortunately, our analyzing must stop there. As stated before, we do not have big enough expected values in any of the other cells to correctly apply the χ^2 test. However, logically analyzing the data, it becomes clear that men seem to swear more than women.

POSSIBLE ERROR

This study is not flawless, however. One error encountered was that my dataset was limited to this corpus. Although the corpus was large, it contained conversations only from 1990 to 1997. Thus, swear frequencies might have changed since the conversations were recorded. Furthermore, when analyzing the average age of the person who swore in my data, the numbers seemed a bit high despite my being extremely conservative in age definition (if I was given a range of a speaker's age, I chose the earliest age possible in the range). The average age of a man who swore was still 38.5, which was a bit higher than what I was hoping it to be. This entails that I most likely am not able to extrapolate my findings to my generation, which is what I was hoping to do. Furthermore, it is my belief that there is a drastic decrease in swear usage from the teenage years to the 30's. Thus, I would predict that if the study were only a sample of college age students, we would find overall swear frequency much higher for each gender.

Possibly the biggest source of bias in my data is within each conversation, misrepresentation of true swear frequency. When speaking with friends, or anybody that you are familiar with, swears seem more prevalent. When speaking with somebody that you are not too familiar with, however, you tend to want to present yourself formally to make a good impression. Use of swears is not generally considered socially acceptable when talking to someone for the first time; they give a 'bad impression.' Consequently, because these conversations tended to happen between two strangers, the overall sample's frequency of swears is most likely lower than the true population's swear frequency. To correct for this, I would need a corpus that has conversations between both acquaintances and strangers so that I would have a more accurate representation of true swear frequency.

FURTHER EXPLORATION

Keeping these errors in mind, we do seem to see a much higher swear rate in males than we do in females. To partially account for these errors, I hope to further adjust my code, and analyze the swear frequencies uttered when one is speaking with the opposite gender as opposed to when the two genders in the conversation are the same. This is of interest because, as stated, people tend to swear less when trying to present themselves formally; they want to be proper when speaking with the opposite sex. I hope to see a difference between swear frequencies when men are speaking with men as opposed to when men are speaking with women. I hypothesize that men are more comfortable with other men, and therefore would be more partial to swearing. The opposite should hold true for women-that is- I expect them to swear more when speaking with women than they do when they speak with men. This was a subproject external to my initial hypothesis. Thus, the data are not included.