Linguistic Self-Expression on Dating Websites

Elizabeth Dellea
Kelly Sadwin
John Barr, Faculty Advisor

May 12, 2016

Abstract

Past linguistic studies and common knowledge state that men and women use language differently. Using sample text from the profiles of OkCupid users as a corpus, these notions are evaluated by comparing linguistic features and generating weighted word clouds. Statistically significant differences were found between users of different sexual orientations and gentations, but not genders. The weighted word clouds tended to show cultural trends among groups rather than differences in the use of language.

1 Introduction

It is commonly believed that men and women use language differently. This assumption appears everywhere from linguistic studies to online articles encouraging women to modify their language so that they may contend with their male peers [4]. A 2000 study on performative gender in online communications claims females are guilty of making more references to emotion, providing more personal information, using more intensive adverbs, and making more self-derogatory comments while males more frequently insult their conversation partners [2].

In this paper, such claims are investigated using natural language processing techniques.

A large sample of text, marked by gender, is needed to evaluate linguistic differences in a population. The text should resemble natural speech and language use as closely as possible, ideally as a form of self-expression. Given these considerations, the most genuine use of text available online comes from social media. However, the corpus of text should not be restricted to the social networks of the researchers. Of the popular social websites, the most serviceable model for finding a sample population that meets these constraints is that of the dating website.

The following section discusses the collection and organization of the corpus of text for study. Sections 3 and 4 detail the methods by which the corpus was analyzed and the results of this analysis. Conclusions and potential areas for future work are found in section 5.

2 Data Collection

The data used for this research comes from the dating website `OkCupid.com`. This website was chosen because it is free to use and the profiles are available publicly. The profiles on OkCupid come complete with demographic information.

2.1 Bot Creation

To gather profiles, two bot accounts were made on OkCupid. The purpose of the bots was to accrue a list of users who appear as "matches", whose profiles could then be pulled for analysis.

The bots were designed to pull as large a pool of matches as possible. Both bots were listed as bisexual and seeking users of all genders. One bot was male, while the other was female. The bots requested to see all users regardless of location. OkCupid computes compatibility using a series of diagnostic questions, of which the bots answered the minimum number.[1]

2.2 Web Scraping

Having set up the accounts, the next part of the process was to download user profiles to the database. This was done in two stages using a custom-built Python web scraper with the BeautifulSoup module.

To use a Python script with the bots, the header of the HTTP requests must be mocked. Because OkCupid logs in its users with cookies, the login cookies were copied directly into the header. In addition, the default header attributes from Python were overwritten by those from a Google Chrome header. Dating websites are prone to bot usage, so the authors took this precaution to avoid detection. In addition to the header, the bots placed a delay between requests to mimic human website usage. These headers proved sufficient to both log in as the bot account and to deflect suspicion from the bot's activities.

In the first stage of web scraping, the bots gathered a list of usernames from the matches page. The Python script repeatedly requested the URL of the OkCupid page at which a user may view their matches. On each refresh, the script pulled the usernames, checked them against a master list of names, and eliminated duplicates, also being sure not to record the usernames of the authors' bots. The bots attempted to gather 1000 usernames each.

Each profile from this master list was downloaded in the second stage of the web scraping process. An OkCupid profile has up to ten sections of text with different prompts for the user to answer. These sections of text served as the subject material on which to perform analysis. In addition to the text,

[1] Due to an oversight, the bots only viewed matches between the ages of 18 and 30.

the scraper also pulled user demographic info, specifically age, location, sexual orientation, gender, and gentation.[2]

The process of downloading pages hit few, if any, obstacles. Within a few hours, over 1700 profiles had been downloaded and were ready for processing.

2.3 Data Storage

A table was created in a SQLite database to store the obtained profiles, with columns for each demographic tag and each of the 10 text fields.

There were many users who had not filled out the text portion of their profiles and could not be used for linguistic analysis. In addition, because the bots' settings allowed them to view matches from anywhere in the world, the database contained profiles in non-English languages. After stripping the database of these profiles, the number of viable collected profiles settled at 1558.[3]

3 Statistical Analysis of Linguistic Features

A popular approach to linguistic analyses seen in past studies is to determine quantifiable linguistic features and compare the metrics across groups of data. This style of analysis was the first approach used in this research project. Seven major features were selected for analysis and computed in Python with the help of the natural language processing module TextBlob.

- Word count
- Unique word count
- Average word length
- Average sentence length
- Percentage of words that are adjectives or adverbs
- Polarity[4]
- Subjectivity[5]

[2] Gentation is a portmanteau of gender and orientation used in OkCupid's source code that refers to the gender(s) of matches that users are listed as "looking for". The three possible options are looking for men, women, or everyone.

[3] This count of 1558 profiles may include a small number of bilingual profiles that contained enough English to avoid detection, though the most egregious were discovered and individually eliminated when creating word clouds (see section 5).

[4] A sentiment analysis feature of the TextBlob module calculates the overall positivity of a text sample as a float between -1 (negative) and 1 (positive).

[5] Also a feature of the TextBlob module. Determines a float between 0 (objective) and 1 (subjective).

After computing these features and storing the results in the database, the numbers were computed along the demographic lines between genders (men vs. women), sexual orientations (straight vs. not straight), and gentations (looking for men vs. looking for women). By dividing the data into two opposing sets, statistical t-tests could be performed to determine a statistically significant deviation between two groups.

The results of these tests suggest the following possibilities:

- Non-straight users have longer profiles than straight users. (99.9% confidence)
- Non-straight users use a wider variety of words than straight users. (99.9% confidence)
- Straight users use more positive language than non-straight users. (99.9% confidence)
- Straight users write more subjectively than non-straight users. (99.8% confidence)
- Users seeking men use longer words than users seeking women. (95% confidence)
- Users seeking women write more subjectively than users seeking men. (95% confidence)

The statistics computed for average sentence length were not included due to the wild variation in sentence length exhibited by users. At the time of this writing, it is the current cultural zeitgeist to type with minimal punctuation. This style inflates the number of words per sentence.[6]

The authors find it interesting that there were no statistically significant differences between men and women[7].

4 Word Clouds

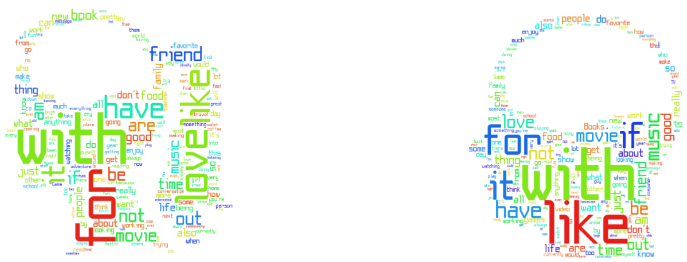
In an attempt to discover further linguistic trends within the data, a visual approach was taken. To see the most frequently-used words within the corpus for each demographic group, word clouds were implemented. As is standard, words with higher frequencies get larger font sizes. This was done with the assistance of the wordcloud Python module.

The wordcloud module by default uses a list of common words that carry little linguistic meaning, or stopwords, including pronouns, articles, prepositions, and conjunctions. Any words in this list are eliminated from the generated

[6] In other words, although the numbers express 99.9% confidence, as humans, the authors have 0% confidence.

[7] This corroborates a 1990 study by Mulac, Studley, and Blau, which found no difference in length between genders.[3]

word cloud to focus on words with more inherent meaning. In initial clouds, the default list of stopwords was edited to remove filler words such as “just” and “feel”, so that these words remained in the word clouds. This was done to investigate the hypothesis that women use these words more often than men.



(a) Female heterosexual users.

(b) Male heterosexual users.

Figure 1: Clouds generated under the original weighting.

Unfortunately, these clouds had extremely similar content to one another using the traditional scheme of weighting word frequencies, as visible in Figure 1. Beyond the striking similarity between the clouds, the expected filler words were not present in the initial female cloud.

To evaluate the relative significance of words used, the second generation of word clouds used a tf-idf weighting scheme. Using tf-idf, each word is assigned a weight that corresponds to both the frequency of the word in a document and the word’s relative uniqueness across all documents.[8] Rather than processing each profile as an individual document, profiles were grouped by the relevant demographics. For instance, when making a cloud for women, profiles were grouped into three documents: male, female, and non-binary genders. The tf-idf clouds for documents based on gender are found in Figures 2, 3, and 4. The tf-idf clouds for documents based on orientation are found in Figures 5 and 6.

The major differences among clouds of gender and orientation are rooted in interests and values, rather than linguistic trends. These findings align with results from a 2001 study on gender in computer-mediated communication (CMC) [5], as well as a survey on CMC studies from 1989-2013 [1].

5 Conclusions and Future Work

Due to the subjective nature of the material studied, the authors hesitate to draw definitive conclusions based on the current body of results. To reiterate, generating word clouds using tf-idf document weighting did not reveal the hypothesized filler words, and analysis of linguistic features showed no statistically

[8] For more information on tf-idf, see `http://lmgtfy.com/?q=tf-idf`.

significant differences between the profiles of men and women, though there were differences found between users of different sexual orientations and gentations.

These results inform further lines of inquiry through which this research could continue. Investigation of the frequency of words seen in the word clouds suggests that the weighting scheme of tf-idf may value the uniqueness of words. Recalibrating the weighting algorithm to strike a different balance between term frequency and inverse document frequency could cause new trends to appear in the word clouds.

The discovery of statistically significant differences between demographic groups could be expanded to include more features for study and more groups to compare.[9] Comparing the data among age groups and locations with the current dataset is flawed due to the limited age range gathered by the OkCupid bots as well as the mutual nature of the matching algorithm. While the bots may be open to all users, the users in turn are more likely to desire matches that are closer to them in age and geographical location. This skews the database towards Ithacans in their early twenties. There is potential in re-scraping OkCupid for profiles from a wider range of ages and locations. In addition, data could be pulled from other source material to sample from a wider population, provided it is also tagged with the appropriate demographics in a reliable way.

Minority gender and orientation identities are not represented prominently in the dataset. In a smaller-scale pursuit of realistic data, OkCupid could be re-scraped to gather profiles specifically from these underrepresented groups.

The authors have shared their database of user profile text on GitHub[10] and invite further exploration and research based upon it.

References

[1] S. C. Herring and S. Stoerger. Gender and (a)nonymity in computer-mediated communication. In S. Ehrlich, M. Meyerhoff, and J. Holmes, editors, *The Handbook of Language, Gender, and Sexuality*, pages 567–586. John Wiley & Sons, Inc, 2014.

[2] M. Hills. *You Are What You Type: Language and Gender Deception on the Internet.* PhD thesis, 2000.

[3] A. Mulac, L. B. Studley, and S. Blau. The gender-linked language effect in primary and secondary students' impromptu essays. *Sex Roles*, 23(9):439–470, 1990.

[4] K. Rogers. Um, like, so... you're killing your own career, 2013.

[5] R. Thomson and T. Murachver. Predicting gender from electronic discourse. *British Journal of Social Psychology*, 40(2):193–208, 2001.

[9] The authors, being computer scientists and not linguists, did not feel comfortable expanding the body of linguistic features.

[10] `https://github.com/bethdellea/TheDatingLang`

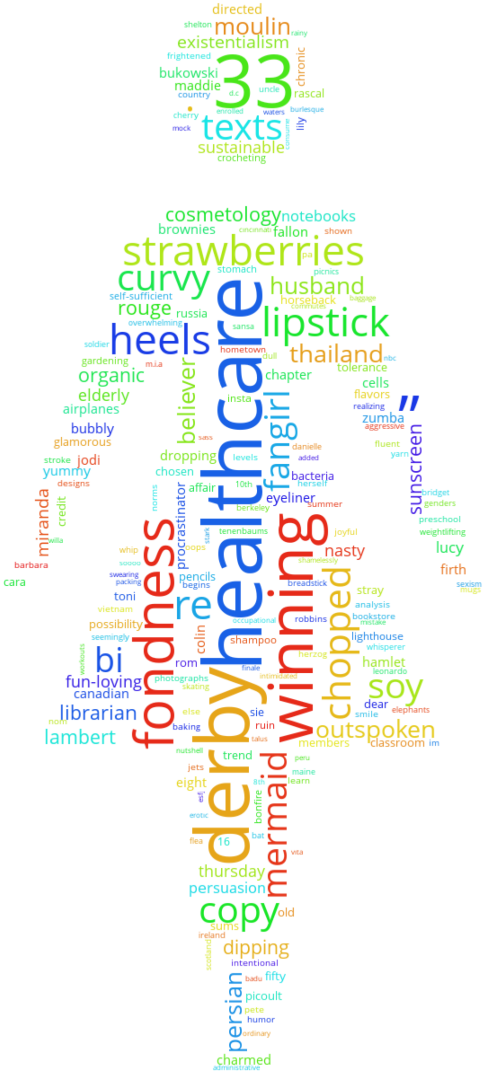


Figure 2: Female user profiles in a tf-idf cloud. Words relating to typically gendered interests appear, such as heels and lipstick.

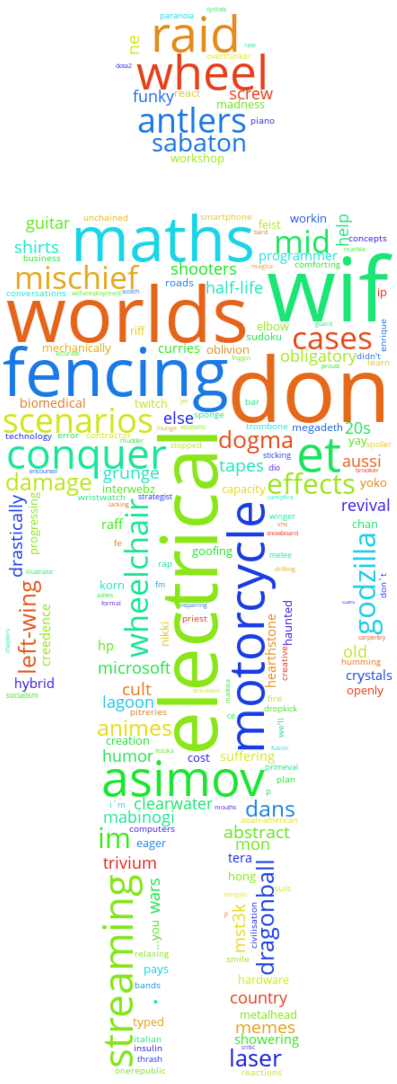


Figure 3: Male user profiles in a tf-idf cloud. Words relating to typically gendered interests appear, such as maths, electrical, and motorcycle.

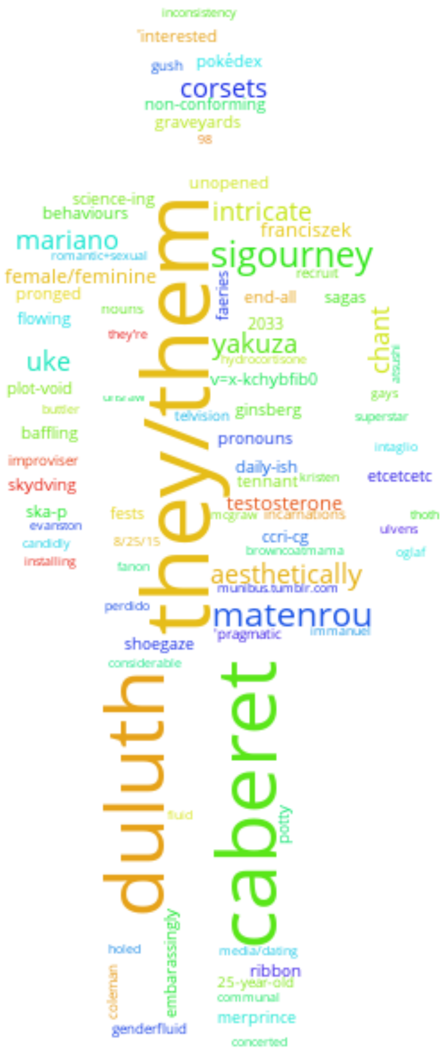


Figure 4: Non-binary user profiles in a tf-idf cloud. They/them pronouns are featured most prominently, as these pronouns are commonly used by non-binary people.

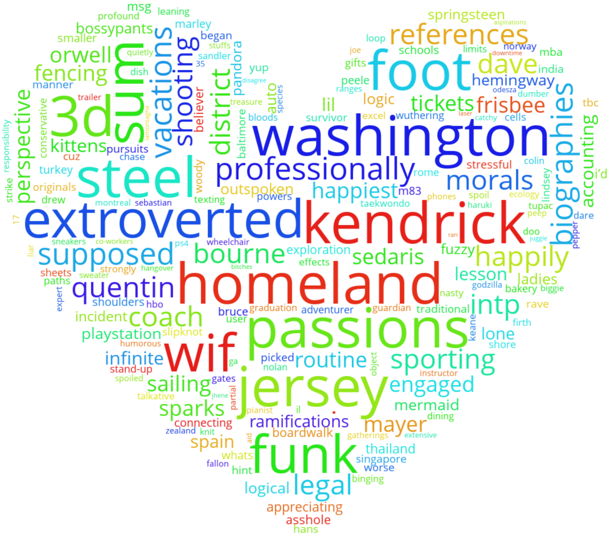


Figure 5: Heterosexual profiles in a tf-idf cloud. Most interesting to observe in contrast with the non-straight cloud.

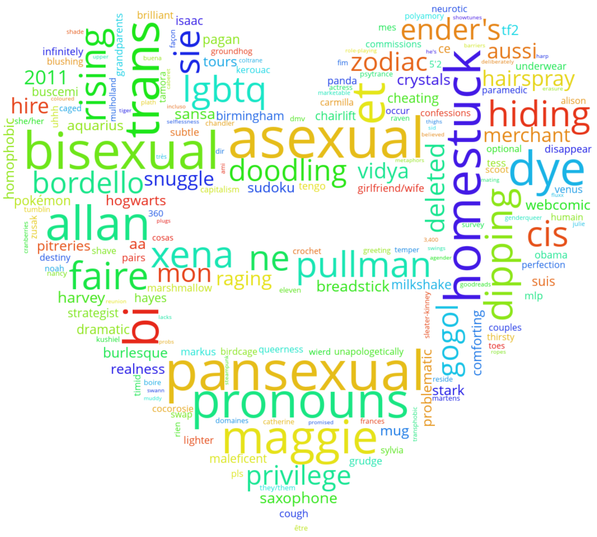


Figure 6: Non-straight user profiles in a tf-idf cloud. Note the orientations and differences in pop culture references from Figure 5.