

An Examination of Language Use in Online Dating Profiles

Meenakshi Nagarajan

Knoesis, Dept. of Computer Science and Engineering
Wright State University, Dayton OH,
meena@knoesis.org

Marti A. Hearst

School of Information
UC Berkeley, Berkeley, CA
hearst@ischool.berkeley.edu

Abstract

This paper contributes to the study of self-presentation in on-line dating systems by performing a factor analysis on the text portions of online profiles. Findings include a similarity in the overall factor structures between male and female profiles, including use of tentative words by men. Contrasts between sexes were also found in a cluster analysis of the profiles using their factor scores. Finally, we also found similarities in frequent words used by the gender groups.

Introduction

Online Dating Systems play a prominent role in the social lives of many people. A few studies have investigated how self-presentation in this setting affects perceived attractiveness (Ellison, Heino, and Gibbs 2006; Hancock, Toma, and Ellison 2007). In one recent study (Fiore et al. 2008), the authors presented online dating profiles to 25 male and 25 female participants, who either saw a component in isolation (the photo, the attributes - age, weight, religion, etc., and the free-text portion), or saw the profile as a whole. They rated the components on attractiveness, trustworthiness, femininity, and masculinity (participants only rated profiles of the opposite sex). As past research in the psychology of attraction would suggest, the photo was the strongest predictor of whole profile attractiveness, but interestingly, the study also suggested that the free-text component played an important role in predicting overall attractiveness.

Men's free-text components were perceived as more attractive when they were also rated as trustworthy, extroverted, and, surprisingly, feminine. This mirrored another finding in the same study that men's photos were more attractive if they were also rated trustworthy, kind, and feminine. For women's text components, attractiveness was not significantly associated with any of the other dimensions. For their photos, femininity was strongly linked with attractiveness, but masculinity was negatively linked.

The fact that women were rating men's photos and free-text components as more attractive when they were perceived as more feminine contradicts standard assumptions (Tannen 1994). One hypothesis is that men may be aware (consciously or not) that self-presenting as more feminine

is a good strategy in online dating, if not in other walks of life. In this work, we extend the study of free-text aspects of profiles on a data set that is an order-of-magnitude larger, to contrast features used by the sexes.

Data and Methodology

Text Corpus: We used data from a popular, paid online dating site - Yahoo Personals. We randomly crawled 500 male and 500 female personals, restricting the reported ages to be between 18 and 60, and restricting to those looking for people of the opposite sex. The free-text component of the profile was the 'Me and my Partner' section where members write about themselves and what they are looking for in a mate. Total number of words in the text portions of male and female-authored corpus was 72049 and 78841 words respectively. Average number of words for the male and female profiles were 134 and 145 words.

Method: We used the features defined by the Linguistic Inquiry Word Count (LIWC)¹, whose dictionary and text analysis program have been used extensively in the past to study links between language and personality traits (Pennebaker and Francis 1996). LIWC covers over 4500 words and word stems spanning linguistic, psychological, personal, paralinguistic and punctuation categories. The LIWC text analysis program analyzes text samples on a word-by-word basis, comparing each word to those in 80 categories. The percentage of total words found in each category is recorded as a variable for every text sample. We obtained such an output for the free-text components of all profiles and conducted a three step analysis as follows:

1. We used exploratory factor analysis to identify the systematic co-occurrence patterns among the LIWC variables in the profiles, following best practices (Costello and Osborne 2005). The resulting factors comprise a set of LIWC variables that tend to co-occur in the profiles. They are interpreted as underlying dimensions of variation based on the assumption that co-occurrence patterns reflect underlying communicative functions. We used the factor analysis tool available with XLSTAT² and used Varimax rotation to facilitate interpretation of factors.

¹www.liwc.net

²<http://www.xlstat.com/en/home/>

Factors, loadings	Male Profiles	Female Profiles
F 1, +ve loadings	Personal and 1st person singular and impersonal pronouns, common, auxiliary and present tense verbs	Personal and 1st person singular pronouns, common, auxiliary and present tense verbs
F 1, -ve loadings	None	Comma (punctuation) and words longer than six letters
F2, +ve loadings	Affect, positive emotion, biological process and sexual words	Affect, positive emotion, biological process and sexual words
F2, -ve loadings	Relativity and space words	Relativity, space and time words
F3, +ve loadings	Impersonal pronouns, cognitive process, tentative and exclusion words	Impersonal pronouns, cognitive process, tentative, discrepancy and exclusion words
F3, -ve loadings	Relative words	Biological process and sexual words

All variable loadings in both datasets were greater than 0.4, with eigenvalues of 3.9, 2.6, 1.6 in the Male profiles and 4.7, 2.9, 2 in the Female profiles. The only crossloading variable was the 'impersonal pronouns' in the Male profiles with a maximum loading of 0.37.

Table 1: Factor structures for Male and Female Personals

2. The obtained factors were interpreted functionally and factor scores for all profiles were calculated using XLSTAT. Profiles have high or low factor scores as their values are high or low on the variables in a factor pattern.

3. While factor analysis delineates dimensions of variation in data, an individual profile may load anywhere on the dimensions. In order to group profiles on the basis of their shared multi-dimensional features, we conducted a k-means cluster analysis using the profile factor scores as predictors. Setting $k = 3$ created the cleanest clusters for both datasets. The profiles grouped into any cluster are intended to be maximally similar in their use of the LIWC variables, while different clusters are maximally distinguished.

Results

Factor Analysis

For the 500 female profiles, a three factor solution involving 20 features turned out to be the strongest. Together the three factors accounted for 48% of the shared variance and were readily interpretable. For the 500 male profiles, a four factor solution comprising 18 features was extracted to be the strongest factor solution. The fourth factor was however not easily interpretable and was discarded. Together the three factors accounted for 45% of the shared variance. Table 1 shows the factor structures obtained for the two datasets. The positive and negative sets of features in a factor occur in complementary distribution, so profiles that have a high frequency of the positively loading features will have a low frequency of the negative set of features.

For the female profiles, the first factor was characterized by positively loading features that relate to immediate interaction and activities. These included use of personal and first person singular pronouns, common, auxiliary and present tense verbs. Negatively loading features included comma (punctuation) and words longer than six letters, both related to elaborate language production. This factor in men's profiles had an additional positive loading of the impersonal pronoun category and no negatively loading factors.

The second factor for both sexes comprised of positive loadings of affect, positive emotion, biological process and sexual words and a negative loading of relativity words

(time, space and motion). The third factor for both sexes was characterized by a positive loading of impersonal pronouns, cognitive process (e.g., think, know), tentative (e.g., perhaps, maybe) and exclusion words (e.g., but, except). Female personals also had a positive loading of the discrepancy word category (e.g., could, should). For the same factor, the male profiles had a negative loading of relativity words while the female profiles had negative loadings of biological process and sexual words.

An important observation was the rate of use of tentative words in male free-text descriptions. 2.6% of all words used by men and 2.2% of all words used by women were classified as tentative words. The mean and standard deviation of tentative word usage in men profiles was 3.5 and 2.4 words, compared to 3.2 and 2.1 words in female profiles. Tentative words are typically attributed to feminine discourse (Tannen 1994), and so this lends further support to the results of (Fiore et al. 2008) described above. It could be that men are temporizing their language and trying to appear attractive in the online profile world.

Cluster Analysis

The k-means cluster analysis based on the three factor scores for each profile uncovered three distinct clusters for each dataset, see Figure 1. The largest cluster for each sex, cluster 1, comprising of 371 male and 363 female profiles, had moderate scores on all three factors. Such an intermediate stance could imply well-balanced profiles that include a mix of all word features. For the two other clusters, there were contrasting patterns between male and female profiles.

In the second pair of clusters, cluster 2 of 70 female profiles, those that focused on immediate interaction and activities, i.e. had high scores on Factor 1, showed moderate use of biological and sexual words and low use of affect, positive emotion words, low use of impersonal pronouns, cognitive process, tentative, discrepancy and exclusion words and a moderate to high use of relativity words and low use.

For the corresponding cluster of 66 male profiles, those that focused on immediate interaction and activities showed high use of affect, positive emotion, sexual and biological words, high use of impersonal pronouns, cognitive process, tentative and exclusion words and fewer relativity words.

In the third pair of clusters, cluster 3, the 67 female profiles that focused less on immediate interaction and activities and more on detailed production, also used more cognitive, tentative, discrepancy and exclusion words, impersonal pronouns and had moderate use of sexual, biological, affect words and fewer relativity words. For the corresponding 63 male personals, profiles that focused less on immediate interaction and activities used fewer cognitive, tentative, exclusion words and impersonal pronouns, fewer sexual, biological and affect words and more relativity words. It is unclear what these bundling of traits mean directly, but the contrast between the sexes in their combinations of usages of words is interesting.

Preliminary Word-level Analysis

The above analysis showed similarities in the word-types that men and women use in self-presentation, and differ-

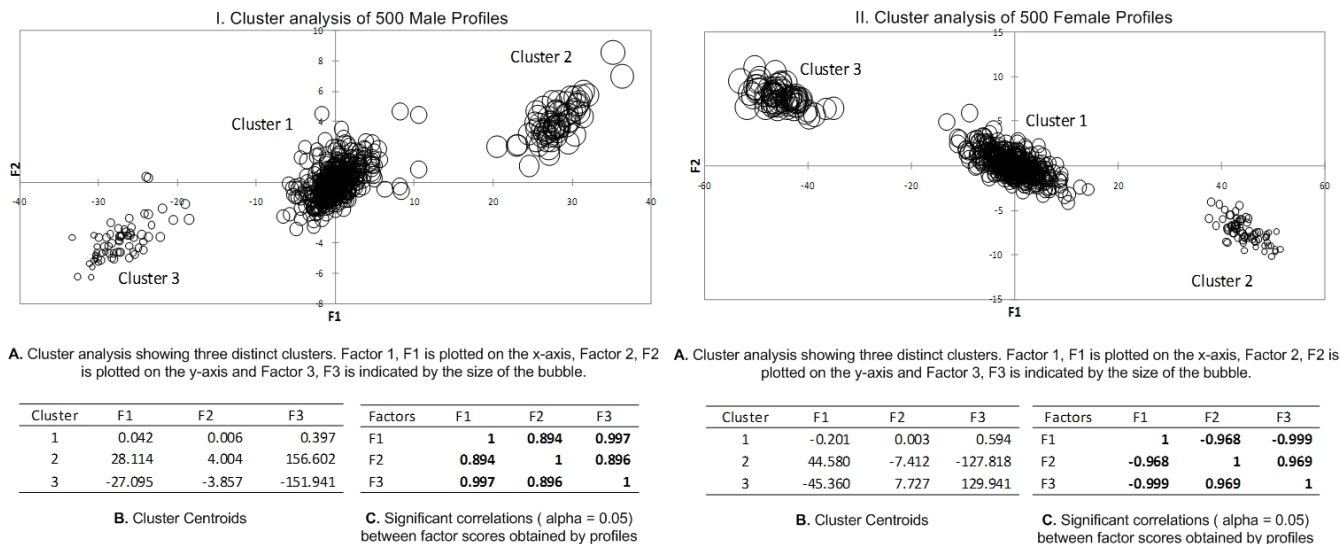


Figure 1: IA and B in the left column show 3 distinct clusters obtained for the male profiles and their centroids, IC shows the correlations between factor scores obtained by male profiles. The column on the right shows corresponding information for female profiles.

ences in how these word types are used together. To cast some light on the word usage, Figure 2 shows in detail the most frequently used words in each LIWC category, grouped by gender. It also shows the proportion of each word used within its category, e.g., for the pronouns category, the most frequently used word was 'my' and it occupied 8% of the pronoun usage in the male corpus and 9% in the female corpus.

For most categories, at least 7 out of the 10 high frequency words were the same for both males and females, and were also used with similar frequencies. We also found that men and women use comparable proportions of LIWC category words in their profiles. Figure 3 shows the percentage of all words used by men and women that fall into each LIWC category. This is calculated by dividing the total number of words used by a gender group from a LIWC category by the total number of words used by that group. Again, the proportions of words that men and women use across these categories is very similar.

Closed-class words (pronouns, prepositions, etc.) are more commonly used than others because they provide the syntactic structure that holds sentences together. However, in past work, gender differences have been seen in their usage. Perhaps more revealing in this data is the similarity in usage of words in the open-class categories as seen in the affect group and the verb groups. It may be the case that self expression tends towards attempting homophily in online dating profiles.

Conclusions

We performed a medium-scale analysis of text in online dating profiles from Yahoo Personals. We found similarity in the overall factor structures between male and female profiles, including the interesting observation of the use of tentative words by men. Clustering the data revealed some sim-

ilarities and some differences in the combinations of word usages between the sexes. Future work should study the impact of these traits on attractiveness perceptions between people represented by the different clusters. A closer look at frequent word usages also revealed some similarities in the words men and women use across different categories. Future work should explore word usage in more detail to identify similarities or differences in how men and women write in online personals.

Acknowledgements

This research was conducted while the first author visited UC Berkeley, and was sponsored in part by NSF DHB-IIS-0624356.

References

- Costello, A., and Osborne, J. 2005. Best practices in exploratory factor analysis: four recommendations for getting the most from your analysis. *Practical Assessment, Research & Evaluation* 10(7):1-9.
- Ellison, N.; Heino, R.; and Gibbs, J. 2006. Managing Impressions Online: Self-Presentation Processes in the Online Dating Environment. *Journal of Computer-Mediated Communication* 11(2):415-441.
- Fiore, A.; Taylor, L.; Mendelsohn, G.; and Hearst, M. 2008. Assessing attractiveness in online dating profiles. *Proceedings of CHI 2008*.
- Hancock, J.; Toma, C.; and Ellison, N. 2007. The truth about lying in online dating profiles. In *Proceedings of CHI 2007*, 449-452. ACM Press New York, NY, USA.
- Pennebaker, J., and Francis, M. 1996. Cognitive, Emotional, and Language Processes in Disclosure. *Cognition & Emotion* 10(6):601-626.
- Tannen, D. 1994. *Gender and Discourse*. Oxford University Press, USA.

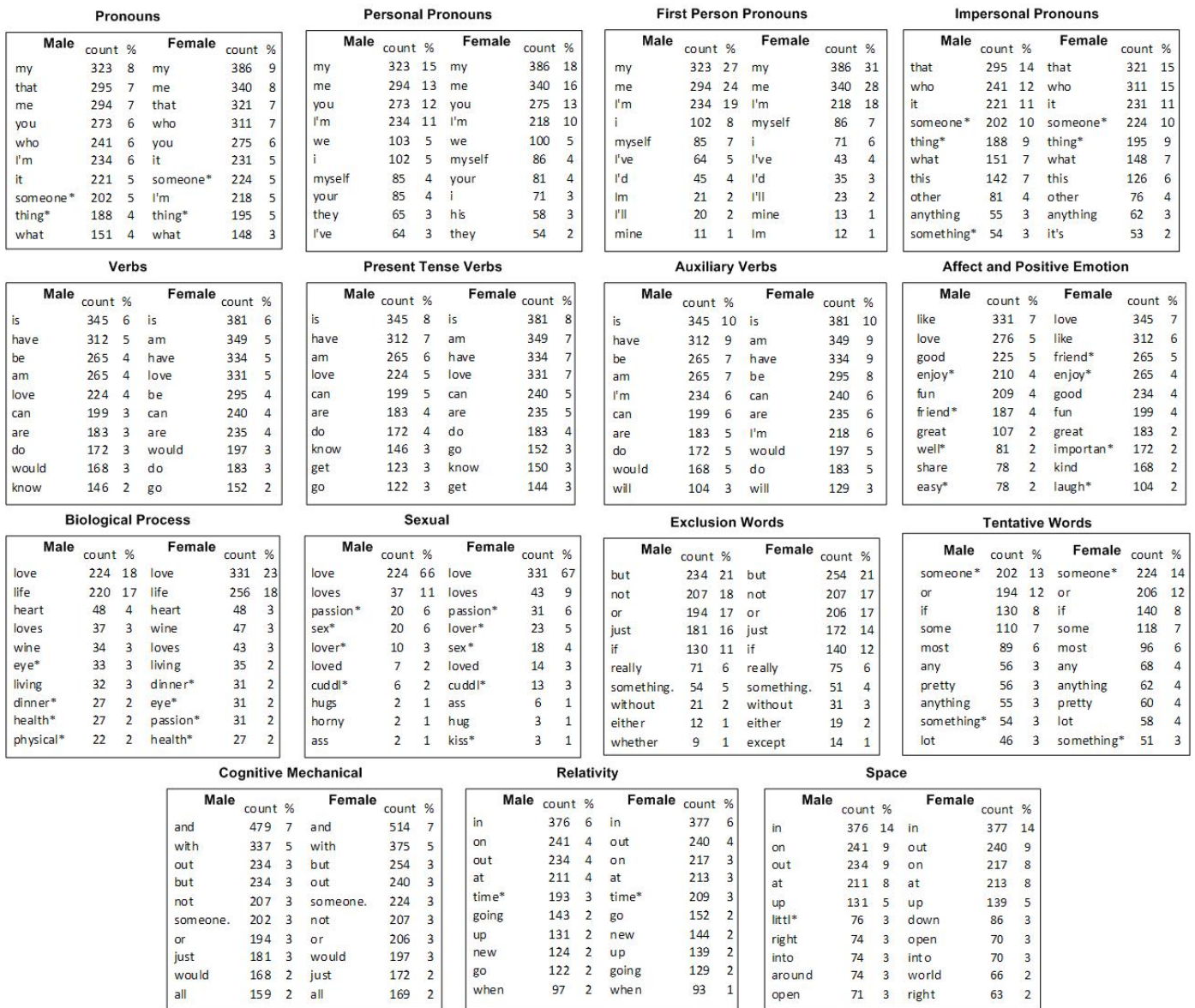


Figure 2: Top 10 frequent words used by men and women for different LIWC categories. Asterisk indicates a word stem that may have several variants (e.g., smok* can be smoke, smoker, smokes, smoking etc.).

Category	Male		Female		Category	Male		Female	
	count	%	count	%		count	%	count	%
Pronouns	11.3	10.6	Biological words	2.3	2.4				
Personal Pronouns	7.7	7.3	Sexual words	0.8	1				
First Person Pronouns	6	5.9	Exclusion words	2.1	1.8				
Impersonal Pronouns	3.6	3.2	Tentative words	2.6	2.2				
Verbs	10.4	9.8	Cognitive Mechanical words	12.9	11.8				
Present Tense Verbs	8.2	7.8	Relativity	9.4	8.1				
Auxilliary Verbs	6.7	6.2	Space	4.4	3.7				
Affect words	8.1	8.1	Time	3.8	3.2				
Positive Emotion words	7.8	7.8							

Figure 3: For each LIWC category, the proportion of words in that category versus all words in the corpus, for both male and female corpora.