Gender Differences across Correlated Corpora: Preliminary Results

Roberta E. Sabin
Computer Science Department
Loyola College
Baltimore, MD 21210
res@loyola.edu

Kerri A. Goodwin
Psychology Department
Loyola College
Baltimore, MD 21210
kgoodwin@loyola.edu

Jade Goldstein-Stewart
U.S. Department of Defense
jadeg@acm.org

Joseph A. Pereira
Computer Science Department
Loyola College
Baltimore, MD 21210
japereira@loyola.edu

Abstract

Does genre affect the way we communicate? We are especially interested in how computer mediation affects the style and content of communication. In an effort to find features related to choice of genre, we have created correlated corpora of writing and speech samples from a single population of subjects. All written and spoken text was opinion-related with topics prescribed and moderated. This paper reports some preliminary results of analysis of a portion of this corpus with an emphasis on linguistic features that may be gender-related.

Introduction

Is it possible to correctly identify the genre of a text sample? Within a given genre, it is possible to correctly identify the identity of the author or, more simply, the author's gender?

As the use of the Internet and the amount of electronic text has grown, interest in the automatic classification of documents has increased. Motivations vary and include, besides author identification, summarization of content, identification of topic, and spam detection. In limited domains classification has been successful: identification of gender of authors within the British National Corpus (Koppel et al. 2002), categorization of news stories and web page descriptions (Calvo et al. 2004, Goldstein-Stewart et al. 2007), differentiation of three authors' works (McCarthy, et al. 2006). In the KDD Cup 2003 Competitive Task, the best system achieved 85% accuracy in identifying scientific articles by the same author when that person authored with over 100 papers (Hill and Provost 2003).

Of course, success at classification rests on the selection and use of linguistic features. Do varying communicative genres have distinct linguistic features? The first comprehensive attempt to answer this question was made by Biber (1988), who selected 67 linguistic features and analyzed samples of 23 spoken and written genres. His results identified six factors that could be used to differentiate different genres of writing.

Since that ground-breaking study, new "cybergenres" have evolved, including email, blogs, chat, spam, and text messaging. Efforts have been made to characterize the linguistic features of these genres (Baron 2005, Crystal 2001, Herring, 1996, Shepherd and Watters 1999, Yates, 1996). The problem is complicated by the great diversity that can be exhibited by even a single genre. Email can be business-related, personal, or spam; the style can be tremendously affected by demographic factors, including gender and age of the sender. Additionally, the context of communication influences language style (Thomson and Murachver 2001, Coupland et al. 1988). Are there patterns that persist for an individual within or across genres? Are these patterns gender-related?

Many studies have attempted to determine "male" and "female" characteristics of communication. More than 30 studies have identified sixteen language gender-related features (Mulac 2001). However, the results may be suspect: many of the studies had very small sample sizes drawn in a non-random way from a non-representative population. When communication genres are limited to computer–mediated communication, conclusions relating to gender differences are few and sometimes contradictory (see Table 1). Gender models have been developed that successfully predict news preference, but these models were based on blog entries whose topics were solely the choice of the author (Liu and Mihalcea 2007). Gender attribution of texts that treat the same topic is more challenging.

The lack of corpora is an impediment to determining common features of communication. Gathering personal communication samples faces privacy and accessibility hurdles. All previous studies, to our knowledge, have focused on one or possibly two cybergenres. To provide additional text samples that may be used for analyzing, comparing and contrasting the communication of individuals and classes of individuals (such as male/female) across different communication modalities, we have created six topic-related or "correlated" corpora. With content limited to opinions on current event topics, we have collected communicative samples from the same individuals on the same six topics in each of six genres: email, essay, phone interview, blog, chat, and in-person, small discussion groups. Here, we discuss the formation of

this corpus and report some preliminary results related to gender identification.

Corpora Collection

Topics and Genres

To ensure the appropriateness of the topics, we conducted a pilot study which resulted in the selection of six topics relevant to student subjects. The topics (the Catholic Church, gay marriage, the Iraq war, marijuana legalization, personal privacy, and gender discrimination) were selected to be controversial and relevant for college students, from whom the subjects would be drawn. We included both conversational and non-conversational genres (Table 2). All genres facilitated the expression individual opinion, but some (blog, chat, discussion) allowed peer give and take.

Participants

In fall, 2006, 24 students were recruited as participants (12 women and 12 men) and balanced the order of presentation of all topics across genres using a Latin Square design. For Phase I (fall, 2006), we collected emails, phone interviews, and essays. In January, 2007, fifteen of the original cohort continued to Phase II. Nine additional students were recruited (of appropriate gender). Phase II (spring, 2007) of the study collected communicative samples via blogs, chat, and in-person small group discussion.

The 45 participating students (pilot study and study) ranged in age from 18 to 29. Of participants in Phases I and II, 69.7% reported their religion as Catholic. All identified English as their primary spoken language. Each participant received a small stipend for participating.

A woman graduate student in psychology served as the phone interviewer and discussion leader. Interviews and discussion groups were all held in the same environment. The graduate student was trained to pose a topical question and to coax participants to continue speaking when a lull in the conversation occurred. She and another research assistant provided the same function in the chat rooms.

Table 1: Gender-Linked Features in Computer-Mediated Communication

Source	Feature	Correlation	Corpus
Savicki, et al (1996)	argumentative words	+males	on-line discussion
Herring (1996)	adversarial language	+males	on-line discussion
Hatt (1998)	Intense adverbs	+ female	1-on-1 chat
	unpleasant passive words	+ males	
Thomson & Murachver (2001)	ref. to emotion	+ female	email
	personal info	+ female	
	modals	+ female	
	intense adverbs	+ female	
Nowson & Oberlander (2006)	pronouns		blogs
	words describing emotional and physical states	+females	
	articles	+males	
	contextuality	+females	
Liu & Mihalcea (2007)	1st person singular pronouns	+females	blogs
	focus on present events	+females	

Table 2: Genres in the Correlated Corpora

Genre	Phase	Computer-mediated	Conversational	Peer Give and Take	Synchronous/ Asynchronous	Audience
Email	I	yes	yes	no	Asynchronous	addressee
Essay	I	no	no	no	Asynchronous	unspecified
Interview	I	no	yes	no	Synchronous	interviewer
Blog	II	yes	no	yes	Asynchronous	world
Chat	II	yes	yes	yes	Synchronous	group
Discussion	II	no	yes	yes	Synchronous	group

Procedure and Design

Each student was asked to express their opinion on each topic in each genre. In each phase of the study using matched random assignment, with gender as the matching variable, two men and two women were randomly assigned to each of the six topic orders. In each phase of the experiment, complete counterbalancing of genre was employed, in which students were randomly assigned to one of six orders of Genre (Phase I: email, essay, and interview; Phase II: blog, chat, and discussion. Transcripts from each session across each type of media and topic were separated into individual files, resulting in 864 text files (several participants produced multiple blog entries). The resultant design was a completely within-participants design, with the exception of replacement participants between Phase I and Phase II of the experiment.

Phase I: Email, Essay, Interview. For emails, participants were given an account on an internal mail server accessible only in a small campus lab. In an effort to control distractions and the influence of non-participants, each participant physically came to the lab, at a time of their choosing, to respond to six email messages from the student research assistant asking their opinion on one of the six topics. For essays, participants were instructed to express their opinions in an essay of approximately 500 words. Students used Word to create the essays which were then deposited in a digital dropbox already familiar to most students. (Note: Although essays were created with computer software, we do not consider them to be computed-mediated communication; students typically use software for creating and transmitting their writing). For interviews, the graduate student interviewed each participant by phone on each of the topics.

Phase II: Blog, Chat, Discussion Group. Students were randomly assigned to a "blog group" of 4 students, 2 men and 2 women. Students selected and used screen names to preserve anonymity. Members of each group blogged on a topic during a two-week period. When sufficient text was acquired (i.e., at least 300 words per participant), the next topic was introduced by the monitoring research assistant. For chat, students were randomly assigned to a "chat group" of 4 students, 2 men and 2 women. A chat room was established on the campus network. As with blogs, each student selected and used a screen name to preserve anonymity. A research assistant moderated each hour-long chat session to keep participants on topic and elicit input from less verbal participants. For each topic, each participant's contributions were extracted to one of four separate files. Students were randomly assigned to a live discussion group of 4 students, 2 men and 2 women. Members of the group met in a comfortable office space and sat at a small table with the moderator, who elicited their interactions on a specific topic. After sufficient text had been acquired from all participants (i.e., approximately 3 to 5 minutes per participant), another topic was introduced. Three topics were discussed per session that ranged in length from 45 to 60 minutes. Discussions were recorded and transcribed, with interviewer input removed. Each participant's contributions were extracted to one of four separate files.

Some General Corpora Characteristics

Word Count

There was no significant difference between males and females in word count across the six genres. This finding is consistent with recent results recently reported by Pennebaker and colleagues (Mehl et al. 2007). There are, however, several interesting patterns illustrated in Figures 1 and 2. Among individual-oriented genres (email, essay, interview), interviews produced a higher mean word count compared to emails and essays. Among the group-oriented genres (blog, chat, and discussion), discussions produced a higher mean word count compared to blog and chat. Both interview and discussion are spoken genres. Also, across topics, males generated significantly larger amounts of text than females on the topic of the legalization of marijuana.

Figure 1: Mean Word Count By Gender

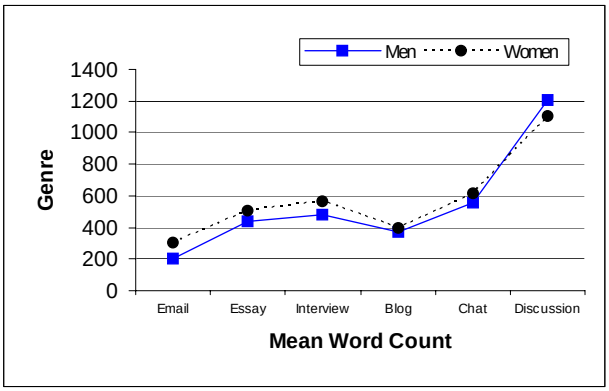
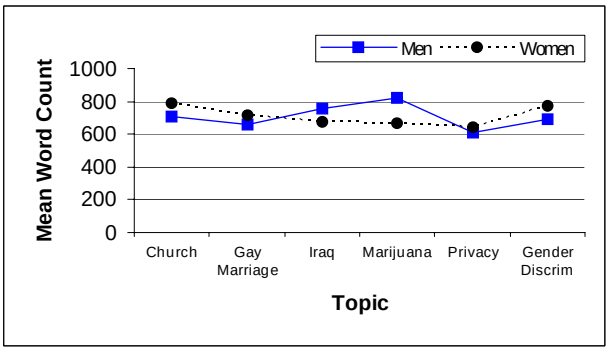


Figure 2: Mean Word Counts for Gender by Topic



Word Frequency

The words most frequently used by males and females were determined and stop words removed. Word lists were generated by genre, by topic, and overall for each

gender. Table 3 shows the words most frequently used by each gender where that word was not as frequently used by participants of the other gender in the given genre (where E=email, S=essay, I=interview, B=blog, C=chat, D=discussion). The table shows words that differed in this way in at least 2 different genres.

Table 3: Frequently used words that were not frequently used by the other gender in the same genre

Females	Genres	Males	Genres
married	E S I B C D	bad	E I D
reason	I B D	guys	C D
children	E S B	public	S I
allowed	E I D	human	E B
student(s)	E S I	issue	B C
woman	E D	fact	S B
couples	E B	problem(s)	B D
officials	E S	pretty	I D

Aggregating all samples from all genres, the top 100 words for males and for females, including stop words, were determined. Seven words differed between male and female in the top 100. The 64 words with counts that varied by 10% or more between male and female usage were selected. Most of these words appeared on the stop word list (www.dcs.gla.ac.uk/idom/ir_resources /linguistic_utils/stop_words). Non-stop word terms included the words “feel”, “catholic” and “school”, which were used more frequently by females then males, as well as the terms “gonna”, “yeah, “yea” and “lot” (used by women) and “say” and “um” (used by men). Some stop words were used more by males ("the", "of"), others by females (“I”, “and”). As this set is mainly stop words, we will refer to this it as the functional word features (F features).

Counts of Words in Word Categories

The frequency of words that belong to word categories was determined using Linguistic Inquiry and Word Count (LIWC). LIWC2001 analyzes text and produce 88 output variables (L features), including some counts of parts of speech but most indicating percentage of total words in given dictionaries (Pennebaker et al. 2001). Default dictionaries were used; these represent categories of words that indicate basic emotional and cognitive dimensions.

Comparisons for gender within each genre were calculated for all 88 features. We selected those features to be potentially gender-related if an F-test performed on the two sets of values (M/F) produced a value < .05 and the feature was present in at least 90% of the text samples. Table 4 shows these selected LIWC features (SL features).

Table 4: Selected LIWC Features in Email, eSsay, Interview, Blog, Chat, Discussion

Category	Examples	E	S	I	B	C	D
Questions	sentences ending ?					F	
Long words	% words > 6 char					M	
All pronouns						F	
1st person sing.	I, me, my		F				
1st person plural	we						F
All 1st person	I, we, me	F					
All 2nd person	you, you'll			M	F		F
Assents	yes, OK				M	F	F
Prepositions	on, to, from	M					
Numbers	one, thirty, million						M
Affective (all)	happy, ugly, bitter	M	F			M	F
Pos. emotions	happy, pretty, good						F
Neg. emotions	hate, worthless		F				
Causation	because, effect				M		
Inhibition	block, constrain				F		
Tentative	maybe, perhaps						M
Certainty	always, never						M
Social processes	talk, us, friend				M		
Communication	talk, share,		M				
Other ref. to people	1st pl, 2nd, 3rd person pronouns		F		M	F	
Family	mom, brother				M		
Time	hour, day, oclock			M	M		
Humans	boy, women					F	
Past tense verbs	walked, were, had		F				
Pres. tense v.	walk, is, be	M					
Fut. tense v.	will, might, shall		F	F			
Space	around, over, up		M				
Motion	walk, move, go				M		
School	class, student,				F		
Leisure	house, TV, music				M		
Home	house, kitchen				M		
Money & finance	cash, taxes				F		
Death	dead, burial,				M		
Body	ache, heart, cough						
Sleep	asleep, bed,				M		
Grooming	wash, bathe, clean				F		
Swear	damn						
Nonfluencies	uh			M			M
All punctuation		F			F	M	

Classification by Gender

Classification of all samples by gender within each genre was performed using four classifiers of the Weka workbench, version 3.4 (Witten 2005). Additionally, the Random Forest classifier (Breiman 2001) was used with 100 trees. Five sets of features were selected: 1. All 88 LIWC features, denoted L; 2. Functional word features only, denoted F; 3. Selected LIWC features, denoted SL (see Table 4); 4. All SL features with the functional word features, denoted SL + F, 5. All LIWC features supplemented by functional word features, denoted L + F.

The best classification results from Naïve-Bayes, J48 (decision tree), SMO (support vector machine) (Platt 1998), Logistic (logistic regression), and RF (Random Forests) are reported in Table 5.

Most of the time, the classifiers were able to correctly identify the gender of the author in a given genre approximately 80% of the time. The one genre that provided an exception was email. Perhaps email messages were too brief, having a lower mean word count (see Figure 1). We plan to further investigate this genre to determine if there are other characteristics, such as the use of the passive voice, that might assist in correctly identifying gender, SMO or RF achieved higher scores than the other classifiers. RF, which is not sensitive to the number of features, always achieved its best scores using the most features (L+F). This makes RF a good general classifier to use with many features. SMO, varied in which features achieved the best score, sometimes by a huge difference.

For the spoken genres, interview and discussion, SMO using F provided the best results. In these genres, LIWC features may need to be supplemented by additional features of the text or features that capture differences in the spoken language word patterns between men and women. It is also possible that the spoken genres, having the highest word counts, biase the word count differences to these two genres. Accordingly, we plan to explore word count statistics within the other genres as well as the use of the entire stop word list to improve accuracy. Using the selected features from LWIC (SL) always

Table 5: Gender classification results for genre and five classifiers. Highest score for the genre is in bold.

Individual Genre	Features	J48	Logs	NB	SMO	RF	Group Genre	Features	J48	Logs	NB	SMO	RF
Email	L+F	59%	60%	63%	67%	**69%**	**Blog**	L+F	62%	53%	67%	76%	71%
	SL+F	54%	57%	59%	60%	64%		SL+F	57%	55%	59%	66%	66%
	F	58%	58%	60%	61%	65%		F	58%	63%	62%	64%	58%
	SL	45%	45%	57%	45%	63%		SL	58%	62%	59%	62%	59%
	L	58%	56%	61%	63%	63%		L	60%	55%	66%	**77%**	69%
Essay	L+F	58%	69%	67%	**80%**	67%	**Chat**	L+F	59%	78%	72%	82%	81%
	SL+F	56%	57%	56%	67%	58%		SL+F	67%	70%	68%	76%	78%
	F	58%	61%	56%	60%	58%		F	63%	61%	66%	75%	76%
	SL	54%	60%	73%	56%	61%		SL	68%	61%	65%	67%	76%
	L	61%	66%	69%	73%	67%		L	65%	70%	72%	**83%**	77%
Interview	L+F	68%	76%	71%	79%	79%	**Discussion**	L+F	67%	78%	83%	81%	83%
	SL+F	59%	72%	74%	85%	74%		SL+F	72%	76%	84%	84%	83%
	F	58%	71%	74%	**86%**	72%		F	63%	77%	81%	**85%**	80%
	SL	43%	45%	65%	49%	60%		SL	69%	70%	68%	69%	67%
	L	64%	68%	59%	73%	67%		L	62%	74%	74%	78%	74%

Figure 3: Classifier Comparison for Gender

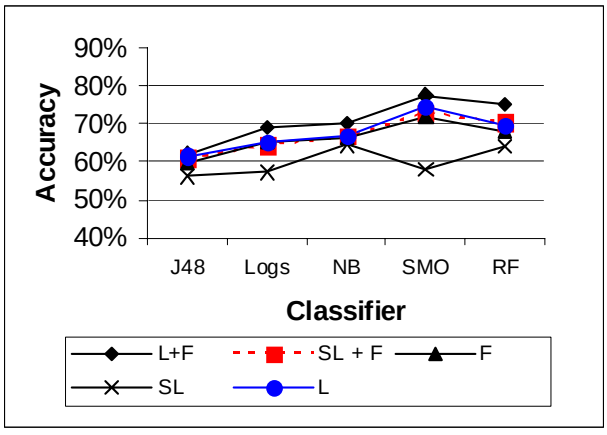
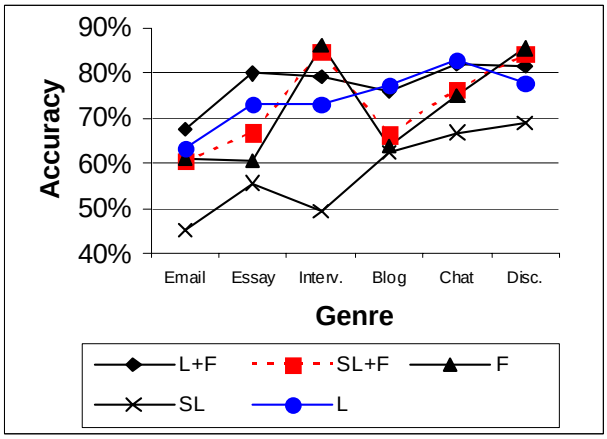


Figure 4: SMO Results for All Genres



resulted in lower performance scores, with the exception of Naïve Bayes classifying essays.

Random Forests and SMO had comparable high performance scores with the exception of the essay genre, where it scored significantly lower than SMO. This bears further investigation.

Conclusion

The results presented here are exploratory and provocative. Besides the determination of additional features that may characterize genres and the author's gender, we will seek to determine their interrelationship and measure their stability. We plan to investigate similarity among text samples across genres from the same subject on the same topic. Additionally, the group conversational corpora (discussion and chat) when separated by subject should yield interesting data for analysis of interpersonal dynamics.

References

Baron, N. S. 2003. Why Email Looks Like Speech. In *New Media Language*, Aitchison, J. and Lewis, D. eds. London: Routledge.

Biber, D. 1988. *Variation across speech and writing*. Cambridge, UK: Cambridge University Press.

Breiman, L. 2001. Random Forests. Technical Report for Version 3, Univ. of California,Berkeley.

Calvo R. A., J. Lee, and X. Li. 2004. Managing Content with Automatic Document Classification. *Journal of Digital Information*. 5 (2).

Coupland, N., et al. 1988. Accommodating the elderly: Invoking and extending a theory, *Language in Society* 17: 1-41.

Crystal, D. 2001. *Language and the Internet*. Cambridge, UK: Cambridge University Press.

Goldstein-Stewart, J, G. Ciany and J. Carbonell, 2007. Genre identification and Goal-Focused Summarization, In *Proc. of the ACM 16th Conference on Information and Knowledge Management (CIKM) 2007*, 889-892.

Herring, S. 2001. Gender and power in online communication. Cen. for Soc. Informatics Work. Paper, WP-01-05.

Herring, S. 1996. Two variants of an electronic message schema. *Computer-Mediated Communication: Linguistic, Social and Cross-Cultural Perspectives*. S. Herring, ed. Amsterdam: John Benjamin. 81-106.

Hill, S. and Provost, F. 2003. The myth of the double-blind review? Author identification using only citations. *SIGKDD Explorations*. 5(2): 179-184.

Klimt, B., and Y. Yang. 2004. Introducing the Enron Corpus. *Proc. Conference on Email and Spam (CEAS 2004)*. http://www.ceas.cc/papers-2004/168.pdf

Koppel, M., S. Argamon, and A.Shimoni. 2002. Automatically Categorizing Written Texts by Author Gender. *Literary and Linguistic Computation*. 17(4): 401-412.

Liu, H., and Mihalcea, R. 2007 Of Men, Women and Computers: Data-Driven Gender Modeling for Improved User Interfaces. In *Proceedings of International Conference on Weblogs and Social Media*. Boulder, CO.

LIWC, Linguistic Inquiry and Word Count. http://www.liwc.net/

McCarthy, P. M., et al. 2006. Analyzing Writing Styles with Coh-Metrix, In *Proceedings of AI Research Society International Conference (FLAIRS)*, 764-769.

Mehl, M. R., et al. 2007. Are Women Really More Talkative Than Men? *Science.* 317, July 2007: 82.

Mulac, A. et al. 2001. Empirical support for the gender-as-culture hypothesis. An intercultural analysis of male/female language differences, *Human Communication Research*. 27: 121-152.

Pennebaker, J. W., M. E. Francis, and R.J. Booth. 2001. *Linguistic Inquiry and Word Count (LIWC): LIWC2001*. Mahwah: Lawrence Erlbaum Associates.

Platt J., 1998. Using Sparseness and Analytic QP to Speed Training of Support Vector Machines. In *Advances in Neural Information Processing Systems 11*, M. S. Kearnset, et al. eds. Cambridge, Mass: MIT Press.

Savicki, V., Lingenfelter, D. & Kelley, M. 1996. Gender language style and group composition in internet discussion groups. *Journal of Computer Mediated Communication*. 2 (3).

Shepherd, M. and Watters, C., 1999. The Functionality Attribute of Cybergenres. In *Proc. of the 32nd Hawaii International Conf. on System Sciences (HICSS1999).*

Thomson, R. and T. Murachver. 2001. Predicting gender from electronic discourse. *British Journal of Social Psychology*. 40: 293-208.

Witten, I. and Frank, E. 2005. *Data Mining: Practical Machine Learning Tools and Techniques (Second Edition).* San Francisco: Morgan Kaufmann.

Yates, S. J. 1996. Oral and Written Linguistic Aspects of Computer Conferencing: A Corpus Based Study. In *Computer-mediated Communication: Linguistic, social, and cross-cultural perspectives.* S. Herring (ed.). Amsterdam: John Benjamin.