Creating and Using a Correlated Corpora to Glean Communicative Commonalities

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

Kerri A. Goodwin
Dept. of Psychology
Loyola College in Maryland
kgoodwin@loyola.edu

Roberta E. Sabin
Dept. of Computer Science
Loyola College in Maryland
res@loyola.edu

Ransom K. Winder
MITRE Corporation
rwinder@mitre.org

Abstract

This paper describes a collection of correlated communicative samples collected from the same individuals across six diverse genres. Three of the genres were computer mediated: email, blog, and chat, and three non-computer-mediated: essay, interview, and discussion. Participants were drawn from a college student population with an equal number of males and females recruited. All communication expressed opinion on six pre-selected, current topics that had been determined to stimulate communication. The experimental design including methods of collection, randomization of scheduling of genre order and topic order is described. Preliminary results for two descriptive metrics, word count and Flesch readability, are presented. Interesting and, in some cases, significant effects were observed across genres by topic and by gender of participant. This corpus will provide a resource to investigate communication styles of individuals across genres, the identification of individuals from correlated data, as well as commonalities and differences across samples that agree in genre, topic, and/or gender of participant.

1. Introduction

Do varying communicative genres have distinct linguistic features? The first comprehensive attempt to answer this question was made by Biber (1988), who selected 67 linguistics 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. A great deal of research has attempted to characterize the linguistic features of these genres (Baron 2005, Crystal 2001, Shepherd and Watters 1999). Motivations vary and include identification of the author, summarization of content, identification of topic, and spam detection. With the exponential growth of cyber communication, the need for automatic processing has escalated. 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 other, demographic factors, including the gender and age of the sender. In addition, it is generally recognized that the context of communication influences language style (Thomson and Murachver 2001, Coupland et al. 1988). So it would be reasonable to assume that, as in other genres, the cyber author alters their style to fit the recipient - one might send very different but topically related emails to a child and to a co-worker. Even if the communication is altered for the intended recipient, there may be common patterns of communication. Samples of written and oral communication for an individual may contain similar content words and patterns of usage. Particular stylistic patterns may persist – characteristics that are unique to a given individual. Recent research has focused on identifying authors within email collections, samples of Reuters news stories, scientific papers, and listserv forums. In the KDD Cup 2003 Competitive Task, the best system was able to successfully identify scientific articles by the same person 45% of the time; for authors with over 100 papers, 85% accuracy was achieved (Hill and Provost 2003).

More globally, in gender identification, there have been numerous studies that attempt to characterize “male” and “female” characteristics of communication. More than 30 studies are summarized in Mulac (2001). Sixteen language features were identified as significantly influenced by gender. However, the results must be suspect: many of the studies cited had very small sample sizes drawn in a non-random way from a non-representative population. Contradictions abound in these studies.

An impediment to determining common features of communication is not computing power but the lack of corpora. To our knowledge, all previous studies have focused on one genre. 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 corpora. Limiting content to the expression of opinion on current event topics, we have collected communicative samples from the same individuals on the same topics in each of six genres: email, essay, phone interview, blog, chat, and in-person small discussion groups. In this paper, we will discuss the formation of this corpus and report some statistics on its composition.

2. Corpus Collection

2.1 Topics and Genres

To ensure that we selected topics that would generate communication samples of sufficient length, we piloted twelve topics as possible suitable topics for the study. These topics were selected to be controversial and were politically and/or socially relevant for college students, from whom the subjects would be drawn. For the pilot study, twelve students conversed with a female interviewer on four of the 12 topics and speaking times were recorded. In addition, the students rated (7 point scale) each topic by comfort level if they were to engage in a conversation on the topic. Based on the rank order of the median speaking times, the normalized speaking times, the variability in speaking times, and the overall student topic comfort level, we selected six topics of for our study (Table 1).

Topic	Question
Church	Do you feel the Catholic Church needs to change its ways to adapt to life in the 21st Century?
Gay Marriage	While some states have legalized gay marriage, others are still opposed to it. Do you think either side is right or wrong?
Privacy Rights	Recently, school officials prevented a school shooting because one of the shooters posted a myspace bulletin. Do you think this was an invasion of privacy?
Legalization of Marijuana	The city of Denver has decided to legalize small amounts of marijuana for persons over 21. How do you feel about this?
War in Iraq	The controversial war in Iraq has made news headlines almost every day since it began. How do you feel about the war?
Gender Discrimination	Do you feel that gender discrimination is still an issue in the present-day United States?

Table 1: Topics

Genre	Phase	Computer-mediated	Conver-sational	Audience
Email	I	yes	yes	addressee
Essay	I	no	no	unspecified
Interview	I	no	yes (speech)	interviewer
Blog	II	yes	no	world
Chat	II	yes	yes	group
Discussion	II	no	yes (speech)	group

Table 2: Genres

We chose to include both conversational and non-conversational genres, hoping to contrast computer-assisted with non-computer-assisted genres (Table 2). The genres email, essay and interview were collected in Phase I and the genres blog, chat and discussion group in Phase II (Sections 2.3.1 and 2.3.2).

2.2 Participants

For our study, we selected 24 students (12 female and 12 male) and balanced the order of presentation of all topics across genres using a Latin Square design. For Phase I, we collected emails, phone interviews, and essays. After Phase I, nine students dropped out of the study; we collected data from nine additional students (4 men and 5 women) to complete Phase II of the study that collected communicative samples via blogs, chat, and in-person small group discussion. In Phase III of the study, we attempted to collect Phase I data (emails, interviews and essays) for the nine “replacement” students added in Phase II. Six students participated, resulting in full samples across the six genres for 21 students. The numbering scheme for these students is displayed in Table 3, as well as which genres these students completed. Within the groupings, participants are sorted by their mean word counts across all genres.

Of the 45 participating students (including the 12 in the pilot study), ages ranged from 18 to 29 years. The majority of participants reported that their primary religion was Catholic (n = 23) and all participants’ primary spoken language was English. All participants received small stipends for their participation.

Participant Number	Genres Completed	Description
P01-P09	E S I	Original Participants, Phase I Only
P10-P24	E S I B C D	Original Participants, Phase I and II
P25-P30	E S I B C D	Replace. Participants, Phase II and III
P31	E B C D	Replace. Participant, Phase II and III (Inc.)
P32-P33	B C D	Replace. Participants, Phase II Only

Table 3: Genres Completed by Participants (P=M for male and F for female). E = Email, S = Essay, I = Interview, B = Blog, C = Chat, D = Discussion.

A psychology woman graduate student served as the interviewer and discussion leader for Phases I and II. She was trained to pose a topical question and to coax participants to continue speaking if and when there were lulls in the conversation. She and another research assistant provided the same function in the chat room setting.

2.3 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 978 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.

2.3.1 Phase I: Email, Essay, Interview

For emails, participants were given an account on an internal mail server accessible only in a campus lab. In an effort to control distractions and the influence of non-participants, each participant physically came several times to the lab, at times 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 that although essays were created by students using computer software, we do not consider these essays to be computed-mediated communication, as most students routinely use such software and frequently transmit their writings via the Internet.)

For interviews, a graduate student interviewed by phone each participant on each of the six topics. The interviewee occupied a faculty office that was modified slightly to be a somewhat more casual setting. Interaction on each topic was of two to nine minutes in length. Interviews were recorded and transcribed, with interviewer input removed.

2.3.2 Phase II: Blog, Chat, Discussion Group

For blogs, students were randomly assigned to a “blog group” of 4 students, 2 men and 2 women. Each student selected and used a screen name to preserve anonymity. Members of the group were instructed to blog on a topic during a two-week period. When sufficient text was acquired (i.e., approximately 300 words per participant), the next topic was introduced by the monitoring research assistant. Blog sites were unprotected but were accessible only on campus. Only study participants could post entries.

For chat room discussions, 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 acted as moderator during 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.

For live discussion groups, students were randomly assigned to a discussion group of 4 students, 2 men and 2 women. Members of the group met in an office space and sat at a small, round table with the moderator, a graduate student 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, and each participant's contributions extracted to one of four separate files.

2.4 Limitations

The correlated corpus is small in size, consisting of approximately 978 text samples,. Additional limitations include the homogeneity of the participants and the fact that the content of the communication was not spontaneous—participants were instructed in written directions or via gentle verbal cues to stay on topic. There were environment constraints: school offices were used for discussion and interview—settings that may prohibit totally free expression (we experienced no swearing).

3. Corpus Analysis

We expect that analysis of the correlated corpus will yield interesting patterns allowing the identification of persistent linguistic features of the genres and possibly individuals. Initially, we contrast the word counts and readability of the samples by gender of the communicant (Figures 1 and 2) and show mean word counts in the corpus (Figures 3-6).

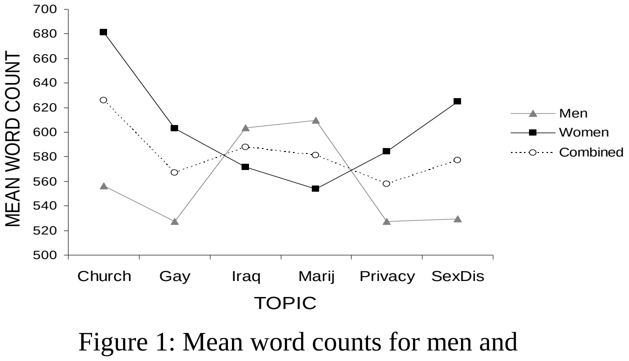

Figure 1: Mean word counts for men and women by topic

3.1 Word Count

In a 2 x 6 x 6 (Gender x Genre x Topic) mixed factorial ANOVA, with Gender as a between-participants factor and Genre and Topic as within-participants factors, was used to assess word counts of the text samples. There was no main effect for gender as women (*M* = 60.597, *SEM* =

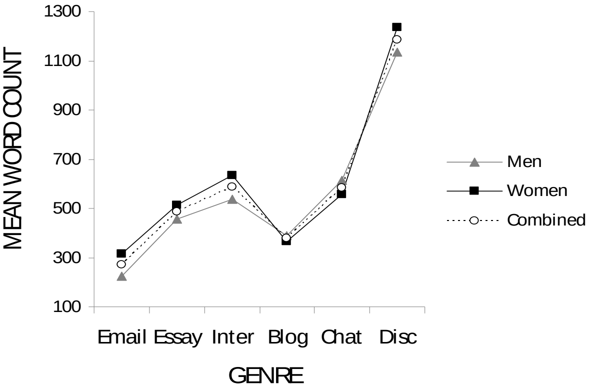

Figure 2: Mean word counts by gender and genre.

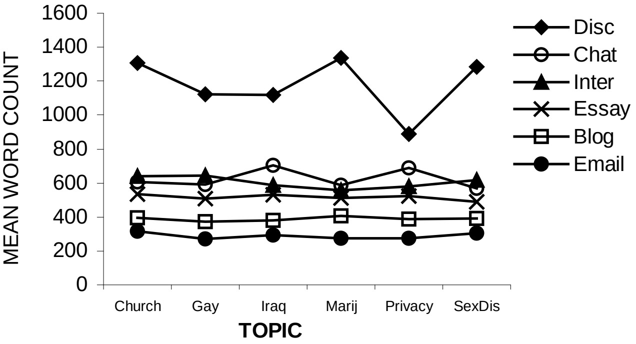

Figure 3: Mean word counts by topic and genre.

42.62) had statistically equivalent word counts as men (*M* = 593.91, *SEM* = 44.70), *F* < 1. There was no main effect for Topic, *F*(5, 95) = 1.75, *p* > .05. However, there was a significant Gender x Topic interaction, *F*(5, 95) = 3.18, *p* < .05 (see Figure 1). Men and women differed substantially in the word counts produced across genres for the differing topics. Specifically, women produced significantly higher word counts than men for the topics of the Catholic Church and sex discrimination, and although it appears that women produced higher word counts for gay marriage and privacy rights, these differences between men and women were not statistically significant. Likewise, word counts between men and women for the topics of the Iraq war and marijuana legalization did not differ.

There was a significant main effect for Genre, *F* (5, 95) = 26.58, *p* < .05 (see Figure 2). Follow-up tests demonstrated that Discussions led to the highest word counts (*M* = 1176.04, *SEM* = 728.75) compared to all other genres. Emails (*M* = 516.62, *SEM* = 19.15), Interviews (*M* = 604.21, *SEM* = 51.60), and Chat (*M* = 624.26, *SEM* = 53.83) led to moderate word counts and these genres did not differ significantly from each other, but all were significantly higher than the word counts for Blogs (*M* = 389.38, *SEM* = 15.64), which was higher than Essays (*M* = 289.12, *SEM* = 32.42). There was also a significant Topic x Genre interaction, *F*(25, 475) = 3.03, *p* < .05 (see Figure 3). The Discussion genre led to the most variability in word counts across topics, particularly for the privacy rights topic. These results indicate the more verbose nature of communication among the Interview, Chat, and Discussion genres that involve .more direct communication with other individuals, particularly for privacy rights topic within the Discussion genre. Figures 4–6 display the variability of mean word count results by individual.

3.2 Readability Measures

Two readability scores were calculated for email and essay genres: the Flesch reading ease score and the Flesch-Kincaid grade level score. The Flesch reading ease score is a rating of text on a 100-point scale, with higher numbers indicated greater ease of readability, and presumably more comprehensible text samples. The Flesch-Kincaid grade level score is a rating of text based on U.S. grade-school level, with scores reflecting the grade-level of the text.

In a 2 x 6 x 6 (Gender x Genre x Topic) mixed factorial ANOVA, there was no significant main effect for Gender and no Gender interactions for Flesch reading ease scores, all *ps* > .05. Women (*M* = 74.24, *SEM* = 1.68) and men (*M* = 70.25, *SEM* = 1.76) did not differ significantly in their reading ease scores across genres and topics.

There was a significant main effect for Genre, *F*(5, 95) = 208.47, *p* < .05. Follow-up analyses indicated that Discussions yielded the highest reading ease scores, followed closely by Interviews, and Chat. Emails, Essays, and Interviews yielded the lowest reading ease scores and did not differ significantly from each other. There was also a main effect for Topic (see Figure 7), in which the Iraq War and Catholic church topics yielded the highest reading ease scores, followed by all remaining topics which did not differ from each other, *F*(5, 95) = 13.23, *p* < .05. More importantly, there was a significant Genre x Topic interaction, *F*(25, 425) = 4.01, *p* < .05 (see Figure 9). For the genres which involved direct interactions with other individuals, Interviews, Discussion, and Chat led to higher reading ease scores than Email, Essay, and Blog with no differences in variability across topics. Since long words affect this score, it was not designed for speech genres which contain disfluencies such as “uh” and “um”, or the abbreviations such as “LOL” that are present in the chat genre. For Emails and Essays, only the topic of the Iraq war led to higher reading ease scores when compared to the other topics in those genres, whereas for Blogs, the topics of marijuana legalization and privacy rights led to lower reading ease scores relative to the four remaining topics.

A 2 x 6 x 6 (Gender x Genre x Topic) factorial ANOVA for Flesch-Kincaid grade level scores also yielded no main effect for Gender and no interactions of Gender with Genre or Topic, p > .05. Women (M = 6.33, SEM = .34) did not differ significantly from men (M = 6.77,

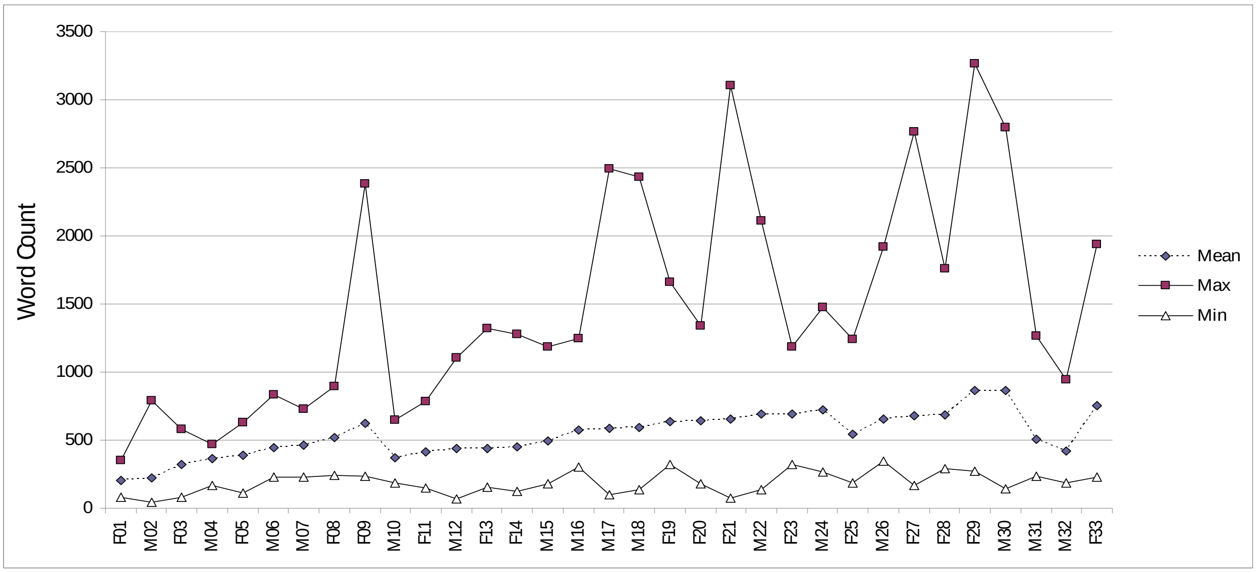

Figure 4: Mean, Min and Max word counts for each individual across topics and genres. Arranged by order in Table 3.

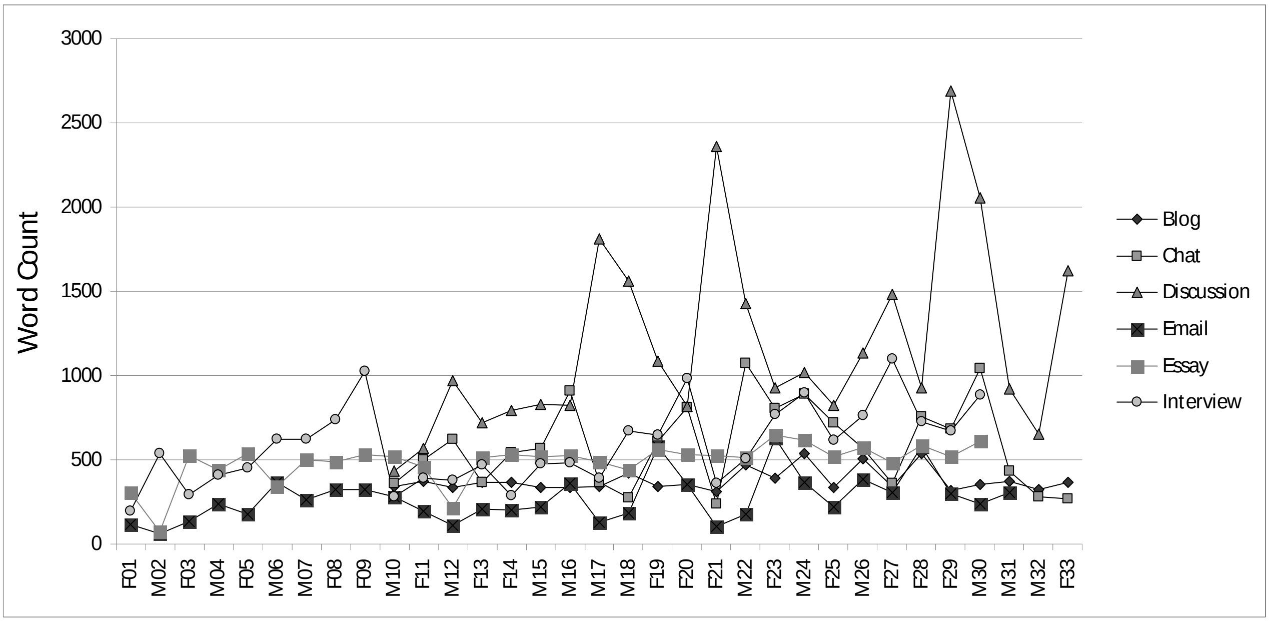

Figure 5: Mean Word Count per Genre for each Individual. Arranged by order in Table 3.

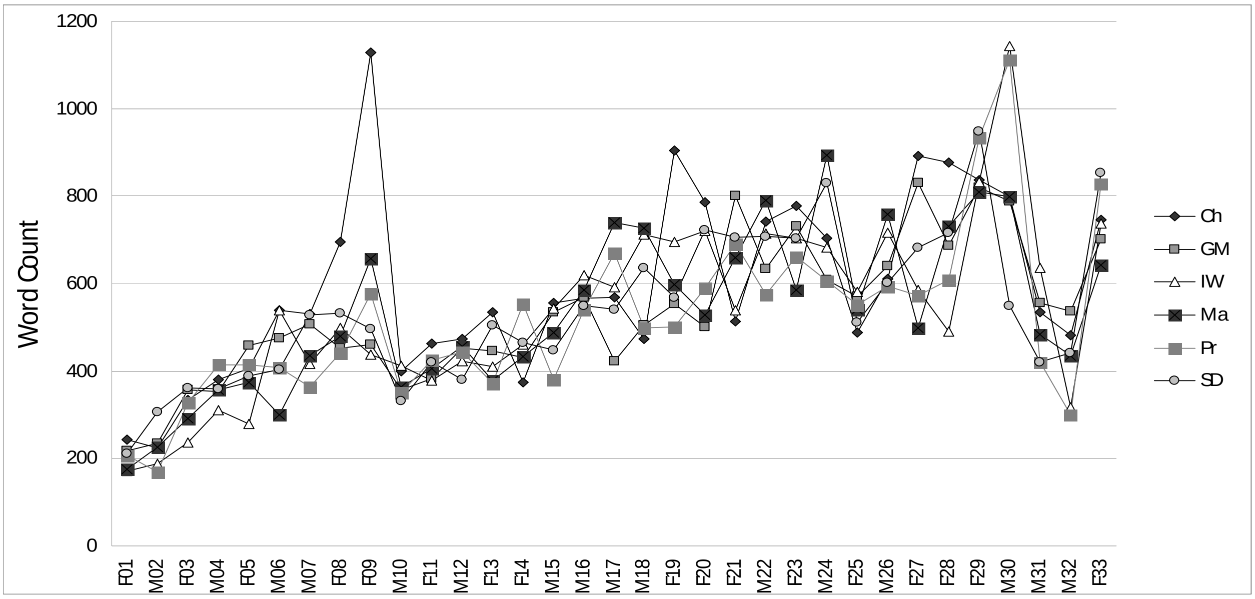

Figure 6: Mean Word Count per Topic for each Individual. Arranged by order in Table 3.

SEM = .36) in their grade-level scores across genres and topics, *F* < 1. There was a significant main effect for Genre, *F*(5, 95) = 219.62, *p* < .05. Follow-up analyses indicated that Blogs led to the highest grade-level scores, followed closely by Essays and Emails. Interviews, Chats and Discussions led to the lowest grade-level scores, and significantly lower than Emails, Essays, and Blogs.

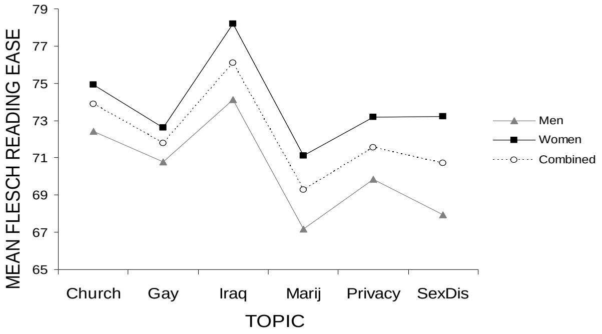

Figure 7: Mean Flesch reading ease scores by topic and gender.

There was also a main effect for Topic, *F*(5, 95) = 8.47, *p* < .05 (see Figure 8). Here, only the topic of the Iraq war led to significantly lower grade-level scores compared to all other topics (which did not differ significantly from one another). There was also a significant Genre x Topic interaction, *F*(25, 475) = 219.62, *p* < .05 (see Figure 10). Interestingly, the results of the Genre x Topic interaction are a mirror image of the reading ease score results, with the Iraq war leading to lower grade-level scores, but only for Email, Essay, and Blog genres.

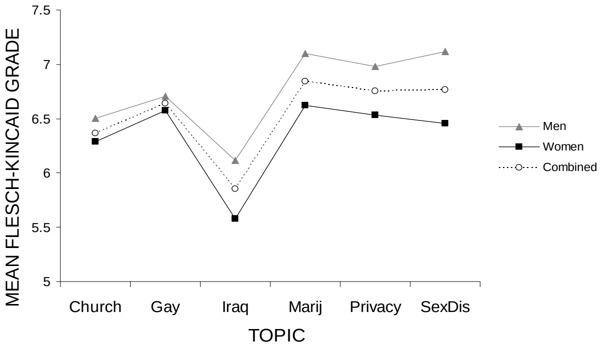

Figure 8: Mean Flesch-Kincaid grades by topic and gender.

It is interesting to note that although reading ease (and presumably comprehensibility) of Interviews, Chats, and Discussions is quite high (Figure 9), the grade-level readability is quite low (Figure 10). On the other hand, the opposite pattern occurs for Emails, Essays, and Blogs. This may be due to the shared ease of communication with others in the more communicative genres (i.e., Interview, Chat, and Discussion), but when put into readability matched for grade-level performance, the coherence of a text that is parsed due to interactions among multiple participants is lost.

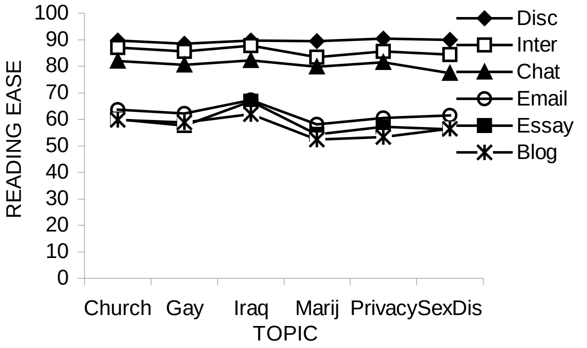

Figure 9: Flesch-Kincaid Reading Ease by topic and genre.

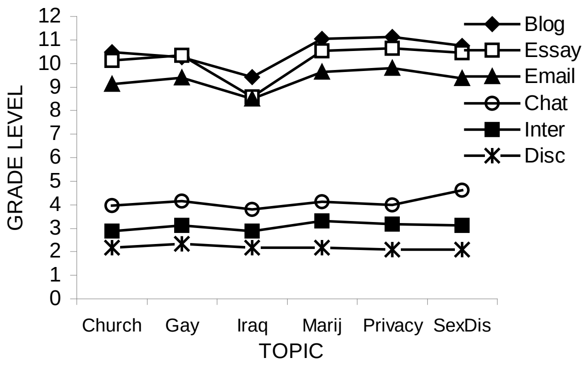

Figure 10: Grade level by topic and genre.

4. Conclusion

In this paper, we have described a correlated corpora collected in order to examine communication patterns from the same individuals across six genres (email, essay, interview, blog, chat, discussion group) and six topics. Although our data is homogeneous in that it represents undergraduate university students and is somewhat constrained , since the data was collected in a prescribed manner and, at times, in a laboratory setting, we believe that the research design allowed us to control two variables: diverse demographics of the subject group and topical content of the communication. Examining word count and readability, we found interesting differences across genres and between the genders.

This corpus will provide additional opportunities to study gender differences within genres and similarities of expression within a genre. It may also allow the discovery of consistencies within communicative samples of an individual across genres that may assist in identification of authorship.

5. Corpus Availability

This corpus will be available to the community in Fall 2008. Please contact the authors or visit http://www.cs.loyola.edu/~res/.

6. References

Banko, M. and Brill, E. (2001). Mitigating the Paucity of Data Problem: Exploring the Effect of Training Corpus Size on Classifier Performance for NLP. In: *Proc. of the Conference on Human Language Technology*.

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

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

Collot, M., and Belmore, N. (1996). Electronic Language: A New Variety of English. *Computer-Mediated Communication*, Herring, S. C. (ed.). Amsterdam: John Benjamins.

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

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

de Vel, O., Anderson, A., Corney, M., and Moha, G. (2001). Mining Email Content for Author Identification Forensics. *SIGNOD: Speicial Section on Data Mining for Intrusion Detection and Threat Analysis*.

Erikson, T. (2000). Making Sense of Computer-Mediated Communication (CMC): Conversations as Genres. *Proceedings of Hawaiian International Conference on System Services (HICSS2000*).

Hale,C. (1996). *Wired Style: Principles of English Usage in he Digital Age*. San Francisco, CA: HardWired.

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

Klimit, B. and Yang, Y. (2004). The Enron Corpus: A New Dataset for Email Classification Research. *Proceedings of the European Conference on Machine Learning (ECML).*

Madigan, D, Genkin, A., Lewis, D., Argamon, S., Dmitriy, F., and Ye, L. (2005). Author Identification on the Large Scale. *Proceedings of the CSNA & INTERFACE Annual Meetings*.

McCombe, N. (2002). Methods of Author Identification. B.A. (Mod) CSLL Final Years Project.

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(1), pp. 121-152.

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

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