In press (2006). *Journal of Sociolinguistics*.

# Gender and Genre Variation in Weblogs

Susan C. Herring and John C. Paolillo
*Indiana University, Bloomington*

*Abstract*

A relationship among language, gender, and discourse genre has previously been observed in informal, spoken interaction and formal, written texts. This study investigates the language/gender/genre relationship in weblogs, a popular new mode of computer-mediated communication (CMC). Taking as the dependent variables stylistic features identified in machine learning research and popularized in a Web interface called the Gender Genie, a multivariate analysis was conducted of entries from random weblogs in a sample balanced for author gender and weblog sub-genre (diary or filter). The results show that the diary entries contained more 'female' stylistic features, and the filter entries more 'male' stylistic features, independent of author gender. These findings problematize the characterization of the stylistic features as gendered, and suggest a need for more fine-grained genre analysis in CMC research. At the same time, it is observed that conventional associations of gender with certain spoken and written genres are reproduced in weblogs, along with their societal valuations.

Key words: CMC, gender, genre, style, text classification, variation, weblog

## 1. INTRODUCTION

Since August 2003, a website called the 'Gender Genie' has purported to identify the gender of the author of samples of written text based on language use.[1] The user can paste in a text sample, select one of three genres—fiction, non-fiction, or weblog—and the Gender Genie tabulates weighted frequencies of 'male' and 'female' grammatical function words; the author's gender is assigned according to whichever category scores highest. While the Gender Genie is intended as entertainment, it is based on serious research: a machine learning algorithm developed by computational linguists (Argomon, Koppel, Fine and Shimoni 2003; Koppel, Argomon and Shimoni 2002; henceforth, 'Argomon and Koppel') to identify author gender in literary and non-literary texts.

It is noteworthy that Argomon and Koppel's research and its popular offshoot, the Gender Genie, both take genre into account. That is, the connection between language use (in this case, personal pronouns, determiners, prepositions, quantifiers, conjunctions, etc.) and writer gender (female or male) is mediated by the classification of texts into conventional types. Genres of communication are traditionally defined in terms of shared purpose and common conventions of content and style (Swales 1990). Efforts at automated text genre identification notwithstanding (e.g. Kessler, Nunberg and Schütze 1997), genre identification must still for the most part be made by a human being, in so far as it involves determination of communicative purpose. In Argomon and Koppel's

computational research and the Gender Genie, word frequencies alone are insufficient to predict author gender; some context must be supplied.

The choice of genres is also interesting, especially the inclusion of 'weblog' in the Gender Genie. Weblogs (blogs) were not included in the original machine learning research, perhaps because they have only recently emerged as a recognized genre of computer-mediated communication (CMC) (Herring, Scheidt, Bonus and Wright 2004; Miller and Shepherd 2004). Yet as the fastest-growing CMC genre,[2] weblogs deserve linguistic study, especially if (as the Gender Genie assumes) language use differs in blogs compared to other genres of writing.

This study extends the empirical investigation of the language/gender/genre relationship to weblogs, operationalized as publicly-available websites, typically single authored, in which dated entries are posted in reverse chronological sequence (Herring, Scheidt et al. 2004). Previous CMC research suggests that gender is reflected in blogging practices. Herring, Kouper, Scheidt and Wright (2004) found a relationship between gender of blog author and blog type: Women write more personal journals, while filter-type blogs, albeit a minority overall, are written mostly by men. Other research (reviewed in section 2.3) has identified gender differences in computer-mediated language, including in weblogs. However, the intersection of these two sets of observations has never been explored, although it raises important questions. Specifically, we wondered: Would male and female bloggers appear to write differently, if blog type were held constant? In other words, are observable variations in language use in weblogs due to author gender, or blog genre (or both, and if so, in what proportions, under what conditions)? The answer to this question bears on the analytical assumptions underlying the Gender Genie, and, ultimately, on the usefulness of the machine learning approach to analyzing language and gender.

To analyze the factors that condition linguistic variation in weblogs, we conducted a multivariate analysis of entries from a balanced sample of random weblogs, taking features of gender-preferential style identified by Argomon and Koppel's machine learning algorithm as the dependent variables, and author gender and blog sub-genre (diary or filter) as independent variables. We expected to find gender style correlating with author gender, and possibly with blog entry genre as well, based on previous research. Surprisingly, however, only genre correlations were found. This leads us to reassess the characterization of the stylistic features identified by Argomon and Koppel as gendered, and to argue for the importance of genre in research on gender and computer-mediated communication.

## 2. GENDER, GENRE, AND LINGUISTIC VARIATION

### 2.1. Spoken Discourse

A relationship among gender, language, and topical domain or discourse genre has been posited by sociolinguists for casual spoken interaction since at least the 1980s. 'Gossip' in all-female groups has been defined as talk about certain topics, such as people,

relationships, and internal states (Aries and Johnson 1983; Coates 1989). All-male groups, in contrast, have been observed to talk more than all-female groups about objects (such as cars and computers) and external events, such as politics and sports (Coates 1993; Johnson 1994). Relatedly, Tannen (1990, 1991) claims that women engage more in 'rapport' talk, and men in 'report' talk.

At the same time, gender differences in discourse genres, like other aspects of English language use, are preferences rather than strict dichotomies (Coates 1993). Thus all-male groups sometimes engage in gossip (Cameron 1997), and female 'geeks' may enjoy computer talk (Bucholtz 2002). Cameron and Bucholtz interpret this to mean that gender identity is flexible and locally constructed through discourse, rather than neatly characterizable as 'male' or 'female'. Nonetheless, generally speaking, there are differences between male and female gossip (Pilkington 1998), and in talk about computers by female and male geeks.

### 2.2. Written Discourse

Certain genres of writing are also traditionally associated more closely with one gender than another. As gossip is considered a women's activity, so diary writing has been described as a women's genre (Heilbrun 1988). Conversely, scientific writing has been claimed to be a masculine genre, traditionally produced by men, and associated with masculine values such as rationality and objectivity (Tillery 2005). These two genres also represent stylistic extremes, with diary writing stereotypically personal and emotional, and scientific writing impersonal and 'plain'. One might ask, however (along with Spender, 1989) whether it is "the writing or the sex" that conditions this stylistic variation, and the societal valuation of men's writing as more serious than women's writing.

Empirical research has investigated the interaction of discourse genre and gendered language in writing. In an experimental study, Janssen and Murachver (2004) asked undergraduate students to write passages involving socioemotional content or political debate. They found that

> [p]assage topic played the greatest role in language use. More female-preferential devices featured in passages involving socioemotional descriptions and more male-preferential features were employed in passages involving political debate. (Janssen and Murachver 2004: 344)

A similar interaction was observed by Argomon and Koppel (Argamon et al. 2003; Koppel, Argomon and Shimoni 2002) in their machine learning research involving a large corpus of written texts from the British National Corpus. Their algorithm identified personal pronouns (favored by females) and noun determiners (favored by males) as significant indicators of author gender. At the same time, they also found 'a strong correlation between the characteristics of male (female) writing and those of nonfiction (fiction)' (321). This led the researchers to construct different statistical models for gender categorization: one for fiction, another for non-fiction. Within each

category, the accuracy of the model in predicting author gender for unknown texts was approximately 80 percent, but accuracy dropped sharply if genre was ignored.

These studies suggest that both gender and genre influence written language, and that some genres exhibit properties traditionally associated with female or male language use.

### 2.3. Computer-Mediated Discourse

Despite claims that the relative anonymity of text-based communication on the Internet would break down traditional gender binaries (e.g. Danet 1998; Rodino 1997), a growing body of research has identified gender differences in computer-mediated discourse, similar to those previously observed in spoken discourse (see Herring 2003, 2004 for overviews). These include a tendency for women to be more polite, supportive, emotionally expressive, and less verbose than men in online public forums. Conversely, men are more likely to insult, challenge, express sarcasm, use profanity, and send long messages. Discussion groups dominated by males have also been observed to use more impersonal, fact-oriented language (Savicki 1996).

As regards topic or discourse genre, Herring (1996a) noted that academic discussion lists that attract widespread female participation tend to focus on 'women's' topics, such as women's studies, women's spirituality, and feminized professions such as education and library science. Discussion lists about computing, in contrast, are overwhelmingly frequented by men, and males predominate in online discussions about politics, philosophy, and linguistics. Herring (1996b) also identified a 'list effect' in academic discussion lists, according to which participants on female-majority lists tend to employ female stylistic features, and participants on male-majority lists tend to employ male stylistic features, regardless of their gender. This recalls the gender/genre interactions reported in the previous sections.

Several studies have characterized weblogs as a new genre of computer-mediated communication (Herring, Scheidt et al. 2004; Miller and Shepherd 2004; Nowson, Oberlander and Gill 2005). The characteristics that make blogs a genre include common structural features such as dated entries displayed in reverse chronological sequence and sidebars containing links and calendars (Herring, Scheidt et al. 2004), a culturally-recognized name (cf. Swales 1990), and the common purpose of sharing content with others through the Web. As with academic discussion lists, however, considerable variation exists according to purpose of communication in weblogs, and a number of researchers have problematized the notion that blogs constitute a unified genre according to the criterion of purpose (Herring, Scheidt et al. 2004; Karlsson 2006; Miller and Shepherd 2004).

In two empirical studies, Herring and colleagues (Herring, Kouper et al. 2004; Herring, Scheidt et al. 2004) performed a content analysis of 357 random weblogs and identified three sub-genres according to blog purpose: personal journals, filters, and k(nowledge)-logs,[3] the first of which is most common. While blog authors were roughly

evenly divided between women and men, gender was skewed in relation to blog sub-genre. More females than males wrote personal journal blogs, although many males wrote journals, as well. However, almost all filters and k-logs were written by men.[4]

Although these two studies took blog sub-genre into account in relation to author gender, they did not investigate the language of blog entries. Other studies have focused on language use in blogs in relation to gender, albeit without making systematic distinctions of blog sub-genre. Nowson, Oberlander and Gill (2005) found gender to be the most significant predictor of variation in a study investigating the effects of individual differences on the formality of writing in a convenience sample of 'personal weblogs'. The writing in women's blogs was significantly less formal than in men's blogs, as measured by Heylighen and Dewaele's (2002) statistical measure of contextuality/formality, which is based on frequency counts of parts of speech. Overall, Nowson, Oberlander and Gill found that the formality of the blogs in their sample was between that for email messages and biography.

Huffaker and Calvert (2005) also analyzed blog language quantitatively, applying DICTION, a content analysis software package, to a gender-balanced corpus of 70 adolescent weblogs. Most of the blogs were drawn from the weblog hosting sites LiveJournal and Blogspot, which are popular with younger authors. The researchers found mixed gender results: Boys used more language classified by DICTION as active, inflexible, and resolute, as in previous gender and CMC research. However, girls and boys made equal use of passive, cooperative, and accommodating language, unlike in previous research.

In a qualitative study, Kennedy, Robinson and Trammell (2005) hand-coded discourse-pragmatic features of comments posted to 20 popular ('A-list') blogs, 10 authored by men and 10 by women. They found that women's comments were more inclusive and expressive, and men's comments were more assertive, competitive, and instrumental, similar to findings of gender differences in other modes of textual CMC (Herring 2003). Furthermore, Kennedy et al. (2005) observed that the women's blogs in their sample were mostly personal diaries, and the men's blogs were mostly political, consistent with the findings reported by Herring, Kouper et al. (2004) of a gender preference in blog sub-genre. However, because their sample varied by genre as well as by gender, it is difficult to tell whether the differences they observed are due to one, the other, or both variables combined.

### 2.4. Research Questions

The research surveyed in the previous sections suggests that both gender and genre condition language use, and that they do so in ways that sometimes result in overlapping patterns (as when certain stylistic features appear to be associated both with a genre and the writer's gender). Thus our overall research question is whether gender or genre is a stronger predictor of linguistic variation in weblog writing.

As prerequisites to addressing this question, we posit two hypotheses:

H1. Male blog authors write differently from female blog authors.
H2. Authors of diary blogs write differently from authors of filter blogs.

Existing evidence suggests H1 to be true. No research has yet evaluated H2, however, or the relationship of H2 to H1.

## 3. METHODOLOGY

To explore the relation between genre and gender variation in weblog language, we conducted a multivariate analysis of stylistic features in entries posted to random weblogs, controlling for gender and blog entry sub-genre. We adopted this quantitative approach in order to evaluate the applicability of the quantitative claims made in Argomon and Koppel's work (and its popularization, the Gender Genie) to weblogs, and because quantification provides the strongest basis for generalization across large data samples, which are readily available in the case of weblogs. We control for gender and blog sub-genre in order to identify the influence of each on the use of the stylistic features; multivariate analysis indicates the relative strength of each influence.[5] This approach is well-established in variationist sociolinguistic analyses of spoken language (Sankoff and Labov 1979), but as yet has been little applied to analysis of computer-mediated language.

### 3.1. Dependent Variables

The dependent variables in this study are a subset of those identified through the text categorization studies of Argomon and Koppel (Argomon et al. 2003; Koppel, Argomon and Shimoni 2002). This approach to gender variation nominates candidates for linguistic variables that have a strong gender component of variation. In Argomon and Koppel's research, the features are a list of English function words and Part of Speech bigrams. Koppel, Argomon and Shimoni (2002) began with a set of 256 features, and winnowed these down to sixteen that contributed the most to distinguishing genders. Among these are the Part of Speech categories pronouns (more common among female writers) and determiners (more common among male writers), and the specific word forms *she, not, for, with*, *and* (all preferred by female writers), *he, the,* and *of* (preferred by males). Since these are common words, they are relatively easy to count, and are not specific to particular topics.

The features we elected to investigate are all specific word forms.[6] These can be broadly categorized into female preferential and male preferential features, following Argomon and Koppel's model. The female preferential features are all personal pronouns: first person singular (*I, me, my, mine*), first person plural (*we, us, our, ours,* and *'s* in *let's*), third person singular (forms of *she* and *he*, counted separately) and plural (*they, them, their, theirs*). The male preferential forms are the determiners *the* and *a/an*, demonstratives, numbers (*1, 2, 1,000, one, two, thousand, first, second,* etc.), other quantifiers, and the possessive pronoun *its*. Argomon, Koppel and Shimoni (2003) characterized the first set as 'interactional' (relating to people) and the second set as

'informational' (specifying objects and concepts), supporting their gender interpretations with reference to qualitative language and gender research.

### 3.2. Independent Variables

The independent variables in this study are blog author gender and blog sub-genre (hereafter referred to as 'blog genre'). Author gender was determined by examining each blog qualitatively for indications of gender such as first names, nicknames, explicit gender statements (e.g. 'I am a 23-year-old British male'), and gender-indexical language (e.g. wife, boyfriend). Gender of blog author was checked for reliability by multiple coders. Only blog authors who were unambiguously identified as either female or male were included in the present study.

For the purposes of this study, blog genre comprised two categories: diary (personal journal) and filter.[7] Herring and her colleagues previously coded this information on the basis of a blog's overall purpose: whether to report and comment on the author's own life (diary), or on events external to the author (filter). In the present study, for greater precision, these two codes were assigned to individual entries, since a single blog may contain posts of more than one genre. Both authors coded blog entries for genre; there was nearly 100% agreement on the genre classifications.

### 3.3. Data Sample

To construct the corpus for this study, we first assembled a list of single-authored candidate weblogs. The list was drawn from two sources, the first a random sample of 100 weblogs collected using the 'random' feature of the blog tracking site blo.gs in March 2004 (see Herring, Scheidt et al. 2004). The second was a larger, snowball sample of blogs collected by following links out three degrees from four random source blogs, also collected during the spring of 2004 (see Herring, Kouper et al. 2005). We initially attempted to construct a corpus balanced for author gender and blog genre entirely from the random corpus, but there was a scarcity of female-authored filter blogs to select from, so we added 20 blogs from the snowball sample, which included more filters.[8]

From this combined pool of candidate weblogs, half male-authored and half female-authored, we selected a sample of current (as of August 2005) entries having a minimum of 100 words each (excluding quoted material), which we attempted to balance among the genre and gender categories. We selected a minimum of two entries from each weblog; where possible, we obtained two of each genre of entry from a single weblog (i.e. we favored mixed-genre weblogs, so as to control for variation in authorial style). The URLs of candidate blogs were sorted by author gender and blog genre, and ordered alphabetically within each category. Selection proceeded by starting at the top of each category, and considering each blog in turn, until a roughly equal number of viable blogs (active, with long enough entries) had been coded for each category.

The final sample comprises 127 entries drawn from 44 different weblogs. Of the entries, 65 were written by women, and 62 were written by men. Table 1 summarizes the distribution of sample entries according to author gender and entry genre.

Table 1. Weblog data sample.

| Author gender | Entry genre | Number of entries |
|---|---|---|
| Female | Diary | 32 |
| Female | Filter | 33 |
| Male | Diary | 31 |
| Male | Filter | 31 |

The sample included 22,134 words by women and 13,587 words by men, for a total of 35,721 words.

### 3.4. Analytical Method

We automatically counted the frequencies of each of the linguistic features and the total number of words for each entry. The frequencies of the features were then analyzed using logistic regression, where the dependent variable is the proportion of the times a feature occurs in an entry out of the total number of words. Two logistic regressions were conducted: one for female-preferential features, and one for male-preferential features. Main effects for feature, entry genre, and gender were estimated, and interaction effects between feature and genre and feature and gender were systematically investigated and tested for significance, using a Wald test.

We reason that, within each analysis, significant gender-feature interactions would indicate features that show a genuine gender pattern of distribution, such that they could be called gender variables. Features that do not show a significant interaction with gender are not gender variables. Similarly, features that show a significant interaction with entry genre are genre variables, while those that do not are not, at least within the confines of the weblog genres (diary and filter) under study. Hence, in these analyses, it is the significance and direction of the interactions that are the main focus of our attention.

Logistic regression was used to analyze the quantitative data (rather than, for example, factor analysis as used in the sociolinguistic genre studies of Biber 1995) for its appropriateness to our focused (rather than exploratory) study design, and since the measures of stylistic features are proportions (rather than e.g. continuous values on an open-ended scale). The statistical analyses were conducted using R, a free/open source statistical programming language and environment.[9] The analyses produced by R are of the same type as those produced by Varbrul, the logistic regression program most familiar to sociolinguists, the only difference being that R's functions for running and reporting logistic regression results follow the standard practice in applied statistics of reporting parameter values on the logit scale (between minus and positive infinity), rather

than the probability weight scale (between 0 and 1) used by Varbrul. (See Paolillo 2002, chapter 8, for further discussion.)

## 4. RESULTS

In both logistic regression analyses, the main effect for gender was not significant. That is, no significant correlation was found between the stylistic features and gender overall. Genre was significant, however, with diaries favoring female-preferential features, and filters favoring male-preferential features. Feature was also signficant; different features have different overall rates of use. Thus the parameter of gender was dropped from the final statistical models, but genre and feature were retained. We tested the significance of the feature-genre and feature-gender interactions in models including these main effects.

### 4.1. Hypothesized Female-Preferential Features

For the hypothesized female-preferential features, all of the feature-genre interactions were significant, meaning that all of the features have different rates of use in the two genres, even taking into account the main effect for genre. In other words, all of the pronoun features show a differential distribution by genre.

In contrast, some, but not all, of the feature-gender interactions were significant. Specifically, gender interactions for *we, he*, and *you* were significant, but those for *she*, *i*, and *they* were not. Moreover, only the direction of *he* and *we* was significantly positive, meaning that it was favored by female authors. The direction of *you* was negative, meaning that it was favored by male authors. Hence, only two out of six of the hypothesized female-preferential features turn out to be so in this analysis,[10] and one is actually male-preferential.

The statistical model that corresponds to these observations is presented in Table 2. The four columns of this table are *Parameter*, giving the name of each effect in the model, its *Estimate* on the logit scale, the *p* value associated with that parameter, and a significance code (* $p < 0.05$, ** $p < 0.01$, * $p < 0.001$). In some cases, because of the nature of the model, the parameters do not directly represent the categories of interest (i.e. feat1, feat2, etc. represent the contrast between the linguistic features as ordered at the bottom of the Parameter column and the reference category, the sixth feature on the list, *we*). (Note: 'gnr1'=filter; 'gnd1'=female.)

To better explain the statistical model, Figure 1 presents the observed frequencies of each of the features. While none of the features is very frequent, of them, first person singular forms are most frequent, followed by first person plural; the remaining features are of roughly comparable frequency.

Table 2. Logistic regression model for hypothesized female-preferential features.

| *Parameter* | *Estimate* | *p* | *Signif.* |
|---|---|---|---|
| (Intercept) | -4.449 | 0.000 | *** |
| gnr1 | -0.078 | 0.001 | ** |
| feat1 | -0.423 | 0.000 | *** |
| feat2 | -0.147 | 0.023 | * |
| feat3 | 1.219 | 0.000 | *** |
| feat4 | -0.308 | 0.000 | *** |
| feat5 | -0.316 | 0.000 | *** |
| gnr1:feat1 | 0.121 | 0.050 | * |
| gnr1:feat2 | 0.222 | 0.000 | *** |
| gnr1:feat3 | -0.480 | 0.000 | *** |
| gnr1:feat4 | 0.151 | 0.007 | ** |
| gnr1:feat5 | 0.173 | 0.002 | ** |
| You:gnd1 | -0.241 | 0.001 | *** |
| She:gnd1 | -0.051 | 0.481 | |
| I:gnd1 | -0.015 | 0.569 | |
| He:gnd1 | 0.209 | 0.004 | ** |
| They:gnd1 | 0.026 | 0.692 | |
| We:gnd1 | 0.163 | 0.005 | ** |

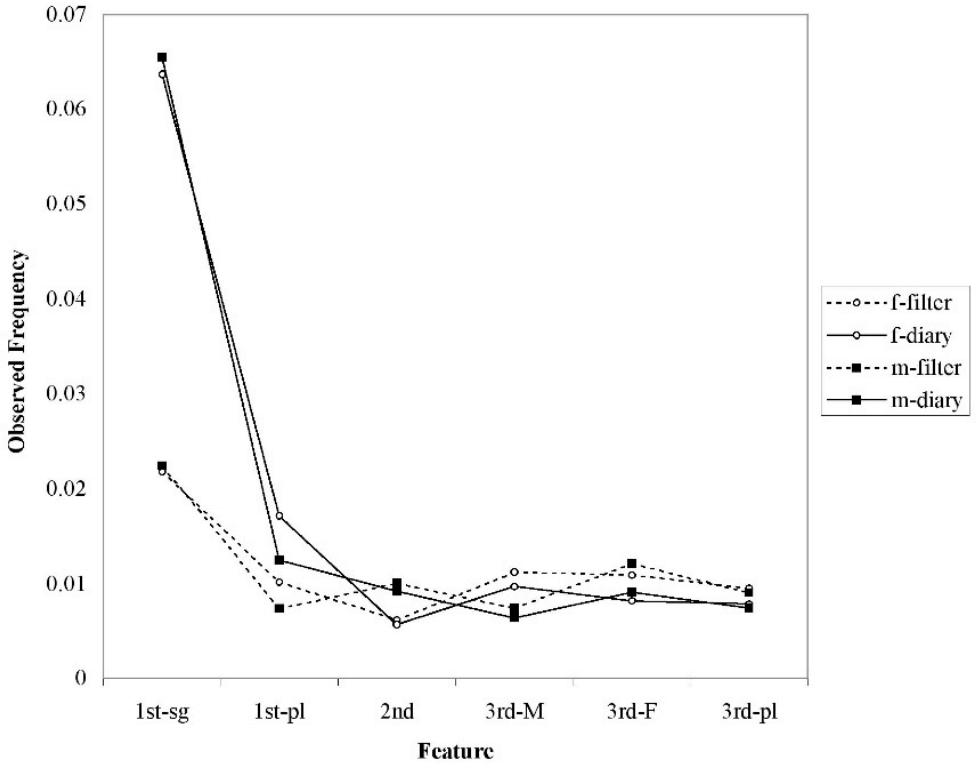


Figure 1. Observed frequencies of hypothesized female-preferential features.

Diary entries show much greater use of first person singular than filters, as one might expect from their different functions. There is very little difference between female- and male-authored entries within either genre. In contrast, first person plural shows clear separation of both gender and genre, with female authors and diaries favoring the use of *we* and its variants *us*, *our*, etc.

Second person shows the unexpected property of being disfavored by female authors, but no genre difference. Third person masculine is favored by female authors, and favored in filters. Finally, both third person feminine and third person plural are favored in filters, but show no significant gender difference.

Hence, in terms of genre, diaries appear to favor first person reference, especially first person singular, as might be expected of personal journals. In contrast, filters favor third person reference, consistent with their focus on external events. Taken together with the main effect for genre and the lack of a main effect for gender, these results show that genre predicts the hypothesized female-preferential features in the weblog entries better than author gender.

4.2. Hypothesized Male-Preferential Features

Table 3 presents the logistic regression analysis of the hypothesized male-preferential features. Fewer of the interaction parameters in this model are significant, meaning that the overall differences in the gender and genre distribution of these features are not as great. (Again, feat1, feat2, etc. represent the contrast between each linguistic feature and the reference category, selected by the model as the last feature in the *Parameters* column, *the*.) (Note: 'gnr1'=filter; 'gnd1'=female.)

The largest effects are the main effects for feature (e.g. *the* is more frequent than *a,* which is more frequent than the other features). Among the interaction effects, only numbers and quantifiers are significant for gender. Of these, moreover, only number is in the expected direction, being favored both in filter entries and by male authors. Quantifiers show an unexpected direction of gender variation, being favored by female authors. Although not significant, demonstratives, quantifiers, *its*, and *the* are also used more often by female than male authors. Most of the features thus do not show evidence of the gender variation we hypothesized based on Argomon and Koppel's work.

Table 3. Logistic regression analysis of hypothesized male-preferential features.

| *Parameter* | *Estimate* | *p* | *signif.* |
|---|---|---|---|
| (Intercept) | -4.338 | 0.000 | *** |
| gnr1 | 0.124 | 0.051 | . |
| feat1 | 0.561 | 0.000 | *** |
| feat2 | -0.177 | 0.024 | * |
| feat3 | 0.199 | 0.006 | ** |
| feat4 | -0.120 | 0.119 | |
| feat5 | -1.884 | 0.000 | *** |
| gnr1:feat1 | -0.077 | 0.271 | |
| gnr1:feat2 | -0.003 | 0.970 | |
| gnr1:feat3 | -0.200 | 0.006 | ** |
| gnr1:feat4 | -0.125 | 0.096 | . |
| gnr1:feat5 | 0.410 | 0.183 | |
| a/an:gnd1 | -0.067 | 0.069 | . |
| dem:gnd1 | 0.045 | 0.415 | |
| num:gnd1 | -0.168 | 0.000 | *** |
| quant:gnd1 | 0.126 | 0.018 | * |
| its:gnd1 | 0.021 | 0.917 | |
| the:gnd1 | 0.042 | 0.087 | . |

Figure 2 shows the observed frequencies of use of each of the hypothesized male-preferential features.

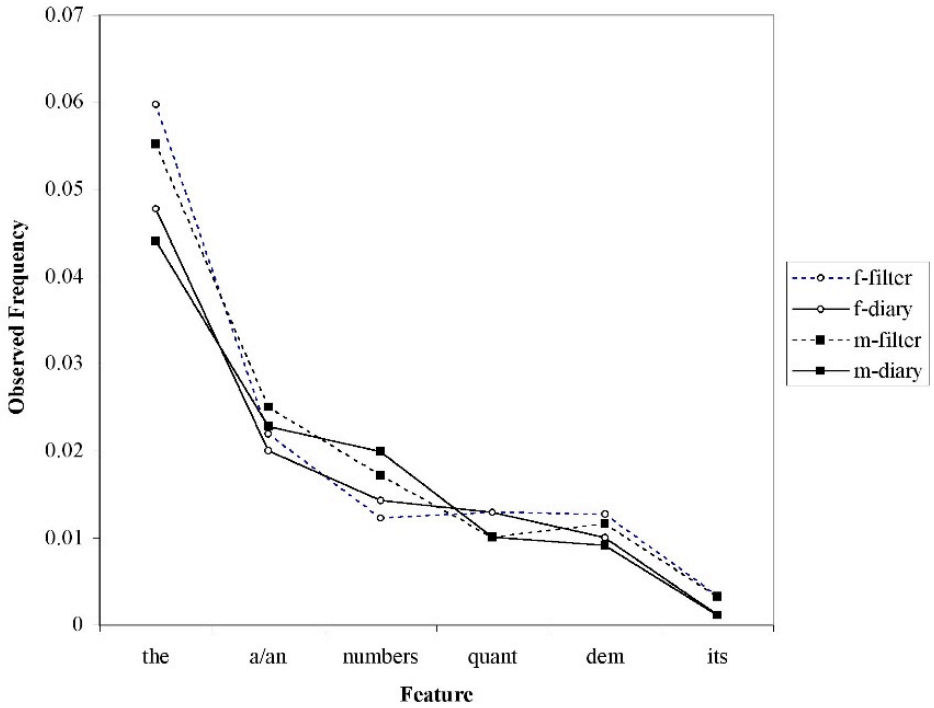


Figure 2. Observed frequencies of hypothesized male-preferential features.

As regards genre, there is an overall positive main effect associating filter genre with the hypothesized male-preferential features. However, only the feature-genre interaction effect for number is significant, and in a negative direction, meaning that diaries use more numbers than filters. Because *the* was taken by the statistical program as the reference category, Table 3 does not include a significance measure for it, but a significant positive correlation with the filter genre can be inferred from the main effect for genre and the observed frequencies of *the* in Figure 2. *Its* is also used more frequently in filter entries, although not significantly moreso than the other male-preferential features.

Hence, among the six hypothesized male-preferential features that we selected for study, only one, numbers, shows the expected association with gender. The genre associations of these features are not as strong as those for the female-preferential features, but filter genre is still an overall better predictor than male gender for this feature set.

4.3. Examples

To illustrate the use of the stylistic features in context, below we present examples of a diary entry (1) and a filter entry (2) from our corpus. The diary entry was written by a male blogger, and the filter entry by a female blogger.

1. Diary entry

> Just a quick note to say that (1) I'm completely rested and recharged, (2) I'm excited about generating high volumes of bloggage and (3) I've seemed to develop a pathological proclivity for prevarication (translation: I'm a big fat liar).
>
> In truth, I've now had three hours of sleep since yesterday at 6:00 a.m., and I'm warily circling this blog like Abbye approaches an operating vacuum cleaner blocking the way to her crate.
>
> I have a load of housekeeping items (in a virtual sense) to get caught up on, not the least of which is finalizing the list of Gazette Blogathon contest winners. In case you're holding off, thinking all prizes have been awarded, think again. No one has yet claimed the grand prize (18 or more header images), there's still one second prize to be awarded, and all three third prizes are open. To refresh your memory on the rules, visit this page.
>
> Now, if you'll excuse me, I think my head has cleared sufficiently so that I can take care of a very important outstanding errand: I have to go shoot a small plastic stool. I shall return after I've exacted my revenge.

Despite being authored by a man, this weblog entry includes more female-preferential words (roughly, nine percent of all words) than male-preferential words (roughly six

percent of words). This pattern is typical for diary entries in the corpus. The most common feature used is *I* and its case variants (14 times out of 194 words).

2. Filter entry

> I have never believed in the Bush concept of pre-emptive war. I think that it goes against everything that American Foreign Policy has ever stood for. Bye-bye Washington's farewell speech, bye-bye Wilson's 14 points, and way to completely negate everything that the Declaration of Independence says about foreign policy. The 2002 Strategy marked a whole new direction in American foreign policy that is frankly a little scary.
>
> The war in Iraq is the first outcome of that Strategy (or the Strategy was written as a way to justify the decision to go to war, which was made in '98). And while I have my own reservations about the how's and why's (not to mention the implications for the concept of sovereignty), there have been some really positive elements from a human rights perspective. Despite the complicated policy issues, knowing that Iraq will never again be tormented by Uday and Qusay Hussein is a huge human rights triumph.

This entry contains more male-preferential features (roughly 10 percent of all words) than female-preferential features (roughly three percent of words), despite being authored by a woman. The most common word is *the* (11 times out of 158 words). This is typical of the filter genre entries in our corpus. Moreover, although we did not measure this, it may be observed that the filter entry makes use of more, and more complex, nominal elements (*Bush concept of pre-emptive war*, *Washington's farewell speech*, *the war in Iraq*, *knowing that Iraq will never again be tormented by Uday and Qusay Hussein*) than does the diary entry in (1), consistent with its focus on concepts and ideas rather than people and processes.

## 5. DISCUSSION

Although we analyzed linguistic features found in previous empirical research to be gender-linked, we found little evidence for systematic gender patterning of the features in a balanced corpus of weblog entries. Author gender did not correlate significantly with either set of features, and when gender-feature interactions were significant, they were nearly as likely to contradict our hypotheses as support them. Thus females favored first person plural (*hypothesized female)*, third person singular masculine (*hypothesized female* in this study; *reported as male* by Argomon and Koppel), and quantifiers (*hypothesized male*), whereas males favored second person (*hypothesized female*) and numbers (*hypothesized male*).

Genre, in contrast, correlated significantly with each set of features (*diary* and *hypothesized female*; *filter* and *hypothesized male*). There was also a larger number of significant genre-feature interactions: Journal entries favor first person singular and first person plural, while filter entries favor third person masculine, feminine, and plural,

determiners, and *its*. The direction of these correlations is not surprising. It makes sense that diaries, which embody a first-person perspective, should use more first person pronouns, and filters, which comment on situations and events external to the writer, should favor third person pronouns and nominal specification. What is surprising is that when blog genre is controlled for, the hypothesized gender differences effectively disappear.

On the one hand, these results provide an unexpectedly straightforward answer to our research question. Genre is a stronger predictor than author gender of the 'gendered' stylistic features identified by Argomon and Koppel (and in the Gender Genie) in our weblog corpus. Indeed, gender is not a significant predictor at all for the stylistic features we investigated.

On the other hand, the results call for explanation, the first and most obvious challenge being to account for why our findings differ from those of Argomon and Koppel. One possible explanation is that our methodology made more fine-grained genre distinctions, classifying weblog content into the sub-genres 'diary' or 'filter'. Argomon and Koppel did not sub-classify genres; rather, they grouped different kinds of texts together under the broad rubrics 'fiction' and 'non-fiction', which are arguably meta-groupings of genres rather than genres per se. (On the question of levels in genre identification, see e.g. Maingueneau 2002 and Swales 1990.) By introducing this distinction, Argomon and Koppel added useful context to the automated text classification approach that resulted in more accurate identification of author gender, but they may not have gone far enough. It is conceivable that further sub-division of their broad genre classifications would have produced results similar to ours, in which stylistic gender differences disappear. If this is so, it suggests that more attention should be paid to genre classification in language and gender research, lest genre effects be mistaken for gender effects.

The second, related, question is why the genres appear to be gendered, when the language of the authors is not. One could argue that the genres we have been discussing are gendered, independent of language. Diary writing has traditionally been associated with females, and politics and external events, the mainstays of filter blogs, have traditionally been masculine topics. Furthermore, previous research shows that females write more diary blogs, and males write a disproportionate number of filter blogs (Herring, Kouper et al. 2004; Kennedy, Robinson and Trammell 2005). But what is the direction of causality, and where does gendered language fit in?

Argomon and Koppel explain their gender findings by invoking the notions 'interactivity' vs. 'informativity'. Females are claimed to be more interpersonally involved and males more informative in their communicative orientation. We suggest that interactivity and informativity are properties of genres, as well (cf. Chafe 1982 on involvement vs. integration as characteristics of speech vs. writing). These properties are best expressed linguistically in certain ways (e.g. first and second person pronouns in interactive discourse; noun specification in informative discourse). This could explain why men and women use similar language within a particular genre; it is the genre's

requirements, rather than the producer's gender, that dictate such usage.[11] The finding in the present study that the interaction effects for features and genre did not exactly match the hypothesized effects for features and gender lends further support to this view.

Finally, it remains to reconcile our findings with a substantial body of previous research on gendered discourse styles in online communication (cf. Herring 2003). Does our finding of no significant gender differences mean that weblogs are more gender-neutral than other modes of CMC, as Huffaker and Calvert (2005) suggest? This is not necessarily the case, for two reasons. First, the fact that Huffaker and Calvert found fewer genre differences than expected in adolescent weblogs may be due to their sample being made up mostly of a single genre of blogs, personal journals. If that is the case, their findings are consistent with those of the present study.

Second, earlier CMC research tended to identify gender differences in terms of discourse-pragmatic usage (such as expressivity, assertiveness, and politeness), rather than in terms of frequencies of grammatical words. Recent research by Kennedy, Robinson and Trammell (2005) found discourse-pragmatic gender differences in comments posted to weblog entries. Based on this, we predict that gender differences in discourse-pragmatic features will also be found in weblog entries, even within the same genre.

At the same time, weblogs make available distinct genres of authorial self-presentation, whose stylistic requirements appear to override some gender-based tendencies in writing, and which are equally accessible to all genders, unlike traditional forms of writing. Moreover, blog genres recall their offline antecedents (e.g. hand-written diaries and letters to the editor; see Herring, Scheidt et al. 2004; Miller and Shepherd 2004) more clearly than do conversational modes of CMC such as chatrooms and discussion forums, in which the conventions for 'written speech' are arguably still emergent. It may be that as more clear-cut genre distinctions emerge in interactive CMC, some apparent gender differences will become increasingly identified with genre rather than with author gender.

Taken together, these observations point to a need to rethink the status of the stylistic variables identified as 'gender' features in Argomon and Koppel's research. We suggest that with reference to weblogs, these might more appropriately be considered genre features.[12]

### 5.1. The 'Gender Genie' Revisited

Like that of Argomon and Koppel, the Gender Genie's treatment of genre is coarse. Only three 'genres' are distinguished: fiction, non-fiction, and weblogs. We have shown that weblogs are not a uniform genre, hence the identification of gender differences in blog posts may be unreliable. In fact, the Gender Genie is only accurate 58.5 percent of the time, according to the site's statistics.[13] Similarly, Cameron (2004) observed the Gender Genie over a three-month period in 2003, and found that it ranged between 51 and 68 percent accurate in its gender attributions. However, Cameron's input texts were weblogs

that came up in a Web search for the term 'Gender Genie'; these may not be typical. The site's reporting statistics could also be misleading for various reasons; for example, users might try to 'game' the system by inputting false reports, or they might be more inclined to report an inaccurate than an accurate result.

To test the Gender Genie's accuracy ourselves, and to consider the possibility that it might predict discourse genre more accurately than it predicts author gender, we pasted in 66 blog entries from our corpus (all entries from every fifth blog, until roughly 16 entries had been assessed from each of the four categories: female-diary, female-filter, male-diary, and male-filter). We selected 'weblog' as the genre in the Gender Genie interface in each instance.

The Gender Genie performed poorly in guessing the gender of the blog authors in our balanced sample; it was correct only 45.5 percent of the time. It was more accurate in predicting blog genre (diary or filter), at 61 percent. The Gender Genie makes use of many of the same features we investigated in this study, albeit weighted according to Koppel, Argomon and Shimoni's (2002) algorithm. Hence the results of this experiment provide additional evidence that the features in question are more strongly genre-preferential than gender-preferential. Neither of the Gender Genie's averages was impressive, however, so the site probably should not change its name to the 'Genre Genie' just yet. It appears that variables in addition to genre influence the language of weblogs; possible candidates for further investigation include topic, register, author stance, and intended audience (see e.g. Scheidt 2006).

## 6. CONCLUSION

In this study of stylistic features claimed to predict author gender, we found genre effects, but no gender effect, in an analysis of entries in random weblogs. This leads us to propose that the functional requirements of the genres investigated—e.g. whether interactive or informative—lead bloggers to employ certain kinds of language, irrespective of their gender. We further propose that a more fine-grained genre analysis of apparently gendered language use in other communicative contexts might also show genre to be a conditioning factor, and that this approach should be pursued in future CMC research.

These proposals are not intended to suggest that gender is irrelevant in weblog writing. Far from it; the fact that the distribution of weblog genres is strongly skewed according to author gender means that, for practical purposes, the language in men's and women's blogs will often differ. Social and political consequences also follow from this distribution: Men's blogs are more likely to appear on 'A-lists' of most popular weblogs (Kennedy, Robinson and Trammell 2005), and to be reported in the mainstream media, in part because filters are considered more informative and newsworthy than personal journals (Herring, Kouper et al. 2004). This recalls the traditional stigma associated with 'gossip' and women's writing (Spender 1989), and reminds us that genres are socially constructed, in part through association with the gender of their producers.

NOTES

[1] The Gender Genie is available at: http://www.bookblog.net/gender/genie.html. Inspired by an article in the *New York Times* about Argomon and Koppel's algorithm, the Gender Genie was programmed for the Web by Rich Miller and Mary Dell, and launched on the BookBlog site on August 15, 2003. The site subsequently became very popular, especially with bloggers.

[2] One of the largest blog-tracking sites, http://blo.gs/, claimed to be tracking over 62.6 million blogs as of February 26, 2006.

[3] Herring, Kouper et al. (2004) and Herring, Scheidt et al. (2004) also coded a 'mixed' blog sub-type (a combination of diary, filter, and/or k-log) and an 'other' type (which in their random samples included poetry blogs, class note blogs, and conversation blogs).

[4] Filter blogs 'filter' content from the rest of the web—often news stories—and link to it, sometimes with commentary attached. Political and issue-focused blogs (such as blogs about the U.S. war in Iraq) are classified as filters. K-logs resemble project logbooks in which notes are kept and information gathered around a particular activity, often technological in nature.

[5] For other methodological approaches to analyzing language and gender in computer-mediated communication, see the references cited in section 2.3, and del Teso Craviotto (this issue).

[6] We did not investigate the part of speech bigrams, as counting those features requires tagging, and a POS tagger would have had to be trained specifically for weblogs.

[7] K-logs were not included in this study because of their low frequency on the public Web. Most k-logs are used in private (e.g., organizational and educational) contexts.

[8] This is a consequence of the sampling method. Following links results in a sample of blogs with links, and filter blogs tend to have more links than journals.

[9] For information on R, see http://www.r-project.org/.

[10] Note that although we included *he* with the other personal pronouns as female-preferential, Argomon et al. (2003) found that male authors used *he* more often. Thus only one out of six of our female-preferential features patterned as predicted by Argomon and Koppel.

[11] This proposal also suggests an explanation for why men and women tend to prefer different genres of discourse and topics of conversation in the first place, i.e. because they are unconsciously mapping their preferences for interactivity or informativity onto generic practices that favor those orientations.

[12] Argomon et al. (2003) also found that genre effects were stronger than gender effects in their research.

[13] See http://www.bookblog.net/gender/stats.php (retrieved January 16, 2006). The accuracy statistics for the Gender Genie are calculated from user feedback; after a

result is presented, the Gender Genie asks 'Am I right?', and the user can click 'yes' or 'no'.

REFERENCES


Argamon, Shlomo, Moshe Koppel, Jonathan Fine and Anat Rachel Shimoni. 2003. Gender, genre, and writing style in formal written texts. *Text* 23 (3): 321-346.

Aries, Elizabeth and Fern L. Johnson. 1983. Close friendship in adulthood: Conversational content between same-sex friends. *Sex Roles* 9 (12): 1183-1196.

Biber, Douglas. 1995. *Dimensions of Register Variation: A Cross-Linguistic Comparison.* Cambridge, U.K.: Cambridge University Press.

Bucholtz, Mary. 2002. Geek feminism. In Sarah Benor, Mary Rose, Devyani Sharma, Julie Sweetland and Qing Zhang (eds.), *Gendered Practices in Language*. Stanford, California: CSLI Publications. 277-307.

Cameron, Deborah. 1997. Performing gender identity: Young men's talk and the construction of heterosexual masculinity. In Sally Johnson and Ulrike Meinhof (eds.), *Language and Masculinity*. Oxford, U.K.: Blackwell. 173-187.

Cameron, Deborah. 2004. Language: Person, gender, number. *Critical Quarterly* 46 (4): 131-135.

Chafe, Wallace. 1982. Integration and involvement in speaking, writing and oral literature. In Deborah Tannen (ed.), *The Oral/Literate Continuum in Discourse.* Norwood, New Jersey: Ablex Publishing Corporation. 35-53.

Coates, Jennifer. 1989. Gossip revisited: Language in all-female groups. In Janet Coates and Deborah Cameron (eds.), *Women in Their Speech Communities*. London: Longman. 94–121.

Coates, Jennifer. 1993. *Women, Men, and Language*, 2nd ed. London: Longman.

Danet, Brenda. 1998. Text as mask: Gender, play and performance on the Internet. In Steven G. Jones (ed.), *Cybersociety 2.0: Computer-Mediated Communication and Community Revisited*. Thousand Oaks, California: Sage. 129-158.

Gender Genie. 2003. Retrieved January 16, 2006 from http://www.bookblog.net/gender/genie.html

Heilbrun, Carolyn. 1988. *Writing a Woman's Life*. New York: Ballantine Books.

Herring, Susan C. 1996a. Posting in a different voice: Gender and ethics in computer-mediated communication. In Charles Ess (ed.), *Philosophical Perspectives on*

*Computer-Mediated Communication*. Albany, New York: SUNY Press. 115-145.

Herring, Susan C. 1996b. Two variants of an electronic message schema. In Susan C. Herring (ed.), *Computer-Mediated Communication: Linguistic, Social and Cross-Cultural Perspectives*. Amsterdam and New York: John Benjamins. 81-108.

Herring, Susan C. 2003. Gender and power in online communication. In Janet Holmes and Miriam Meyerhoff (eds.), *The Handbook of Language and Gender*. Oxford, U.K.: Blackwell Publishers. 202-228.

Herring, Susan C. 2004. Computer-mediated communication and woman's place. In Robin Tolmach Lakoff [Mary Bucholtz (ed.)], *Language and Woman's Place: Text and Commentaries*. New York: Oxford University Press. 216-222.

Herring, Susan C., Inna Kouper, Lois Ann Scheidt and Elijah Wright. 2004. Women and children last: The discursive construction of weblogs. In Laura Gurak, Smiljana Antonijevic, Laurie Johnson, Clancy Ratliff and Jessica Reyman (eds.), *Into the Blogosphere: Rhetoric, Community, and Culture of Weblogs*. Minneapolis, Minnesota: University of Minnesota. Retrieved February 22, 2006 from http://blog.lib.umn.edu/blogosphere/women_and_children.html

Herring, Susan C., Inna Kouper, John Paolillo, Lois Ann Scheidt, Michael Tyworth, Peter Welsch, Elijah Wright and Ning Yu. 2005. Conversations in the blogosphere: An analysis 'from the bottom up'. *Proceedings of the Thirty-Eighth Hawai'i International Conference on System Sciences.* Los Alamitos, California: IEEE Computer Society Press.

Herring, Susan C., Lois Ann Scheidt, Sabrina Bonus and Elijah Wright. 2004. Bridging the gap: A genre analysis of weblogs. *Proceedings of the Thirty-Seventh Hawai'i International Conference on System Sciences*. Los Alamitos, California: IEEE Computer Society Press.

Heylighen, Francis and Jean-Marc Dewaele. 2002. Variation in the contextuality of language: An empirical measure. *Foundations of Science* 6: 293-340.

Huffaker, David A. and Sandra L. Calvert. 2005. Gender, identity, and language use in teenage blogs. *Journal of Computer-Mediated Communication* 10 (2), article 1. Retrieved January 16, 2006 from http://jcmc.indiana.edu/vol10/issue2/huffaker.html

Janssen, Anna and Tamar Murachver. 2004. The relationship between gender and topic in gender-preferential language use. *Written Communication* 21 (4): 344-367.

Johnson, Sally. 1994. A game of two halves? On men, football and gossip. *Journal of Gender Studies* 3 (2): 145-154.

Karlsson, Lena. 2006. Acts of reading diary weblogs. *Human IT* 8 (2): 1–59.

Kennedy, Tracy L. M., Joanna S. Robinson and Kaye Trammell. 2005. Does gender matter? Examining conversations in the blogosphere. Paper presented at Internet Research 6.0: Internet Generations, Chicago, Illinois, October 5-9.

Kessler, Brett, Geoff Nunberg and Hinrich Schütze. 1997. Automatic detection of text genre. *Proceedings of ACL-97*: 32-38.

Koppel, Moshe, Shlomo Argomon and Anat Rachel Shimoni. 2002. Automatically categorizing written texts by author gender. *Literary and Linguistic Computing* 17 (4): 401-412.

Maingueneau, Dominique. 2002. Analysis of an academic genre. *Discourse Studies* 4 (3): 319-342.

Miller, Caroline R. and Dawn Shepherd. 2004. Blogging as social action: A genre analysis of the weblog. In Laura Gurak, Smiljana Antonijevic, Laurie Johnson, Clancy Ratliff and Jessica Reyman (eds.), *Into the Blogosphere: Rhetoric, Community, and Culture of Weblogs*. Minneapolis, Minnesota: University of Minnesota. Retrieved February 22, 2006 from http://blog.lib.umn.edu/blogosphere/blogging_as_social_action_a_genre_analysis_of_the_weblog.html

Nowson, Scott, Jon Oberlander and Alistair J. Gill. 2005. Weblogs, genres, and individual differences. Paper presented at Cogsci 2005, Stressa, Italy, July 21-23.

Paolillo, John C. 2002. *Analyzing Linguistic Variation: Statistical Models and Methods.* Stanford, California: CSLI Publications.

Pilkington, Jane. 1998. 'Don't try and make out that I'm nice!' The different strategies women and men use when gossiping. In Jennifer Coates (ed.), *Language and Gender: A Reader*. Oxford, U.K.: Blackwell. 254-269.

Rodino, Michele. 1997. Breaking out of binaries: Reconceptualizing gender and its relationship to language in computer-mediated communication. *Journal of Computer-Mediated Communication* 3 (3). Retrieved February 26, 2006 from http://jcmc.indiana.edu/vol3/issue3/rodino.html

Sankoff, David and William Labov. 1979. On the uses of variable rules. *Language in Society* 8 (3): 189-222.

Savicki, Victor. 1996. Gender language style and group composition in Internet discussion groups. *Journal of Computer-Mediated Communication* 2 (3). Retrieved January 16, 2006 from http://jcmc.indiana.edu/vol2/issue3/savicki.html

Scheidt, Lois Ann. 2006. Adolescent diary weblogs and the unseen audience. In David Buckingham and Rebekah Willett (eds.), *Digital Generations: Children, Young People and New Media*. London: Lawrence Erlbaum.

Spender, Dale. 1989. *The Writing or the Sex or Why You Don't Have to Read Women's Writing to Know It's No Good.* Athene Series. Oxford, U.K.: Elsevier.

Swales, John. 1990. *Genre Analysis: English in Academic and Research Settings.* Cambridge, U.K.: Cambridge University Press.

Tannen, Deborah. 1990. Gender differences in conversational coherence: Physical alignment and topical cohesion. In Bruce Dorval (ed.), *Conversational Coherence and its Development*. Norwood, New Jersey: Ablex. 167-206.

Tannen, Deborah. 1990. *You Just Don't Understand: Women and Men in Conversation*. New York: Harper Collins.

Tillery, Denise. 2005. The plain style in the seventeenth century: Gender and the history of scientific discourse. *Journal of Technical Writing and Communication* 35 (3): 273-289.