Gender, genre, and writing style in formal written texts

SHLOMO ARGAMON, MOSHE KOPPEL, JONATHAN FINE,
and ANAT RACHEL SHIMONI

Abstract

This article explores differences between male and female writing in a large subset of the British National Corpus covering a range of genres. Several classes of simple lexical and syntactic features that differ substantially according to author gender are identified, both in fiction and in nonfiction documents. In particular, we find significant differences between male- and female-authored documents in the use of pronouns and certain types of noun modifiers: although the total number of nominals used by male and female authors is virtually identical, females use many more pronouns and males use many more noun specifiers. More generally, it is found that even in formal writing, female writing exhibits greater usage of features identified by previous researchers as 'involved' while male writing exhibits greater usage of features which have been identified as 'informational'. Finally, a strong correlation between the characteristics of male (female) writing and those of nonfiction (fiction) is demonstrated.

Keywords: *gender-based writing style; computational stylistics; comparative linguistics; discourse analysis; stylometrics.*

1. Introduction

The question of identifying and interpreting possible differences in linguistic styles between males and females has exercised linguistic researchers for decades (e.g., Trudgill 1972; Lakoff 1975; Labov 1990; Coates 1998). It has been argued for some time that some consistent differences exist in speech (as summarized in Holmes 1993), although the interpretation of such differences remains somewhat elusive. Most previous work has investigated apparent phonological and pragmatic differences between male and female language use in speech (e.g., Trudgill 1972; Key 1975; Holmes 1990;

0165–4888/03/0023–0321 *Text* 23(3) (2003), pp. 321–346

Labov 1990; Eckert 1997) and informal writing (such as student essays (Mulac et al. 1990; Mulac and Lundell 1994) and electronic messaging (Herring 1996)).

Several statistical phenomena have emerged that appear to be fairly stable across a variety of contexts. For example, females seem to talk more about relationships than do males (Aries and Johnson 1983; Tannen 1990) and use more compliments and apologies (Holmes 1988; Holmes 1989) and facilitative tag questions (Holmes 1984). Holmes (1993) has suggested that these and other phenomena might be generalized to a number of 'universals' including the 'universal' that females are more attentive to the affective function of conversation and more prone to use linguistic devices that solidify relationships. However, interpretation of the underlying linguistic phenomena, particularly as regards their specific communicative functions, is the subject of considerable controversy (Bergvall et al. 1996). For example, it has been argued (Cameron et al. 1988) that the use of facilitative tag questions by women might be more plausibly interpreted as signs of conversational control than as signs of subordination, as had been previously contended (Lakoff 1975). Nevertheless, broadly speaking, the differences between female and male language use appear to be centered about the interaction between the linguistic actor and his or her linguistic context (the listener as well as the larger speech community).

Hence it is not surprising that nearly all of the work on male/female linguistic difference has focused on speech and other high-interaction linguistic modalities (such as correspondence). Formal written texts such as books and articles, on the other hand, which are intended for a broad unseen audience, lack the intonational, phonological, and conversational cues that are involved in speech and to a lesser extent in correspondence. One might therefore expect, especially in view of the interactional nature of the differences seen thus far between female and male language use, that such differences would be reduced or even eliminated in such formal written texts. Indeed, some authors (Berryman-Fink and Wilcox 1983; Simkins-Bullock and Wildman 1991) have asserted that no difference at all between male and female writing styles should be expected in more formal contexts.

In this article we explore possible variation between male and female writing styles in Modern English, by studying a large subset of the British National Corpus (BNC) covering a range of different genres. The documents included in this study are all articles and books intended for an unseen audience. Nevertheless, we will identify several classes of simple lexical and syntactic features whose occurrences in texts differ substantially according to author gender, both in fiction and in nonfiction. To foreshadow the main results, we will find significant differences between male- and female-authored documents in the use of personal pronouns and certain

types of noun modifiers: although the total number of nominals used by male and female authors is virtually identical, females use many more pronouns and males use many more noun specifiers.

Our main interest in this article is to present the linguistic phenomena; we will endeavor, as far as possible, to avoid baseless speculation with regard to interpretation of the data. Nevertheless, the differences we consider between male- and female-authored documents represent related underlying phenomena. The categories of pronoun and specifier both encode information about the 'things' of the world as they are presented in nominal groups (Halliday 1994). Pronouns send the message that the identity of the 'thing' involved is known to the reader, while specifiers provide information about 'things' that the writer assumes the reader does not know. Thus, one main locus of difference between men's and women's writing is the way the people, objects, collectives and institutions are presented. In particular, since we will see that it is specifically pronouns that refer to animate 'things' that are used with greater frequency in female-authored documents, our results are consistent with earlier findings that men talk more about objects, while women talk more about relationships (Aries and Johnson 1983; Tannen 1990).

We will see that our results are also consistent with earlier work on relatively small corpora of epistolary writing in the seventeenth and twentieth centuries (Biber et al. 1998; Palander-Collin 1999), in which a difference was found on the 'involvement-informational' dimension (Biber 1995), with women's writing exhibiting more usage of features identified as 'involved' and men's writing exhibiting more usage of features identified as 'informational'. In fact, we will show that for each of a range of individual features, which collectively constitute a good part of the 'involvedness-informational dimension, there are significant differences between male and female usage. The results reported here are particularly surprising because our corpus crosses several genres and thus, unlike a corpus of personal letters, should not be expected to implicate directly the social roles of the writers and the purposes of the texts. Moreover, in the texts examined here, the writers did not have a clear notion of the sex of the intended reader so that any differences in the properties of the writing must reflect characteristics of the writer rather than those of the reader.

Many of the differences we find hold both for fiction and for nonfiction. Interestingly, those features for which there are significant differences between male and female usage also tend to be those for which there are significant differences between nonfiction and fiction. Those features that are more prevalent in male writing are almost invariably more prevalent in nonfiction.

At this stage it is premature to advance strong cognitive speculations as underlying the differences found in the corpus. It may well be that the differences reported here reflect subtle sociological effects that affect perceptions of self and perceptions of the world that are then encoded into the texts.

2. Overview

Studies of gender-based differences in language usage have come under attack in recent years. It has been argued (Bing and Bergvall 1996) that many such studies are methodologically flawed for they assume that significant differences exist and then engage in fishing expeditions to identify them. Mindful of this critique, we have taken great pains to avoid such bias in this study. First, we selected a large, high-quality, genre-controlled corpus as will be described in detail below. Second, we applied fully-automated methods to answer the following question: given a corpus of labeled male- and female-authored documents, can we successfully identify author gender of unseen documents? We found that we could do this with approximately eighty per cent accuracy (Koppel et al. 2001). The bulk of this article will consider the kinds of features which best facilitate the classification of documents by author gender.

3. The corpus

We used a corpus consisting of 604 documents from the British National Corpus (BNC). Each document in the BNC is labeled for genre and all words are tagged for parts of speech from the BNC's tag set of 76 parts of speech (e.g., PRP stands for preposition, NN1 stands for singular noun, and AT0 stands for article and punctuation marks). For each genre we used precisely the same number of male- and female-authored documents (fiction: 123 male documents, 123 female documents; nonfiction: 179 male and female documents respectively, including natural science: 2 documents each; applied. science: 13; social science: 60; world affairs: 34; commerce: 4; arts: 31; belief/thought: 18; leisure: 17). Documents were chosen in each genre by using all available documents in the smaller (male or female) set and randomly discarding the surplus in the larger set. No single author wrote more than six documents in this corpus. All the documents are in Modern (post-1960) British English. The average document length is just above 42,000 words so that the full dataset contains just over 25 million words.[1]

We collected statistics for a set of just over 1000 features that were chosen solely on the basis of their being more or less topic-independent. The features included a list of 467 function words and a list of n-grams of parts of speech (that is, sequences of n consecutive parts of speech appearing in the text) consisting of the 500 most common ordered triples, 100 most common ordered pairs and all 76 single tags. For example, a common triple is: PRP_AT0_NN1 as in the phrase '. . . *above the table* . . .'. Part of speech n-grams were used to more efficiently encode the heavier syntactic information that has previously been shown (Baayen et al. 1996; Stamatatos et al. 2001) to be useful for distinguishing writing styles, in the context of authorship studies.[2]

4. Main distinguishing features

We used a version of the EG algorithm (Kivinen and Warmuth 1997), which is a generalization of the Balanced Winnow algorithm (Littlestone 1987), to automatically select the features that are most useful for properly categorizing a document (Koppel et al. 2001). Briefly, the idea is to use labeled documents in a training corpus to incrementally adjust the 'weight' given to each feature as a male or female indicator: ultimately, some features converge to high male weights, some features converge to high female weights and most features are given little, if any, weight at all. A broad range of machine learning methods such as those we used have proved to be successful at text categorization (Sebastiani 2002). Balanced Winnow, in particular, has been shown to be useful for text categorization and especially for selecting out a small set of features which truly distinguish between corpora (Lewis et al. 1996; Dagan et al. 1997).

The short list of features (less than 50), which our algorithm identified as being most collectively useful for distinguishing male-authored texts from female-authored texts was very suggestive. This list included a large number of determiners (*a, the, that, these*) and quantifiers (*one*, *two, more, some*) as male indicators. Moreover, the parts of speech DT0 (BNC: a determiner which typically occurs either as the first word in a noun phrase or as the head of a noun phrase), AT0 (BNC: a determiner which typically begins a noun phrase but cannot appear as its head), and CRD (cardinal numbers) are all strong male indicators. Conversely, the pronouns (*I, you, she, her, their, myself*, *yourself*, *herself*) are all strong female indicators.

Although a given feature's usefulness for distinguishing male documents from female documents, as determined by Balanced Winnow, does not necessarily reflect the feature's mean frequency difference between males and females, a comparison of male and female usage of pronouns and

Table 1. *Frequency means, medians, and standard errors for pronouns (PNP) and determiners (AT0 or DT0) in male/female and fiction/nonfiction documents*[3]

Feature/Dataset	Female $\mu \pm$ stderr	Male $\mu \pm$ stderr	t-test	Female median	Male median	Mann-Whitney U test
Pronouns/Nonfiction	390 ± 19	282 ± 12	$p<0.000$	1315	242	$p<0.0001$
Pronouns/Fiction	977 ± 18	860 ± 18	$p<0.000$	11016	854	$p<0.0001$
Determiners/Nonfiction	1152 ± 12	1247 ± 8.9	$p<0.000$	11149	1247	$p<0.0001$
Determiners/Fiction	908 ± 13	1041 ± 10	$p<0.000$	1889	1047	$p<0.0001$

determiners (see Table 1) reveals significant differences both for fiction and for nonfiction. These differences are significant both with regard to mean frequencies and median frequencies.

The extent to which frequencies of determiners and pronouns alone can be parlayed into effective categorization of unseen documents as male-authored or female-authored is illustrated by the following fact: of the 59 documents in the corpus where *the* appears with frequency < 0.0524 and *she* appears with frequency > 0.0188, all but two are by females. In fact, as mentioned above, we find overall that unseen documents can be correctly categorized on the basis of features considered in this study with an accuracy of about eighty per cent (Koppel et al. 2001).

From a functional point of view (Halliday 1994), this suggests that different foci characterize the way male and female writers signal to the reader what 'things' are being talked about. The pronouns of women's writing, as all pronouns, present things in a relational way: 'I know that you know what I am referring to, therefore I will present the information as if we both know it'. The specifiers found more frequently in men's writings send the message of: 'here are some details about the things being mentioned'. As we shall see, these differences align with differences between what has been termed 'involved' and 'informative' writing (Biber 1995), as well as with differences between fiction and nonfiction. After considering the statistical differences between male and female writing in some detail, we will consider a number of passages taken from the BNC that illustrate these differences.

5. Female markers: pronouns

Closer analysis of these phenomena revealed several interesting facts that shed further light on this observation. First of all, the extraordinary difference in pronoun frequency between male and female documents does not reflect greater frequency of nominals (common nouns, proper nouns, and

Table 2. *Frequency means for nominal types across sex and genre*

	Fiction		Nonfiction	
Nominal type	Female	Male	Female	Male
Common nouns	1479	1596	2022	2061
Proper nouns	198	226	213	232
Pronouns	978	860	390	282
Total	2655	2682	2625	2575

pronouns, including possessive pronouns) in female documents. In fact, the respective frequencies of nominals in female and male documents (see Table 2) are nearly identical, both in fiction and in nonfiction. Thus there is no discernable difference between males and females in the overall number of references to 'things' in the texts, which emphasizes the prominence of pronouns in female-authored documents.

If we examine relative frequency of pronoun use more deeply (see Tables 3 and 4), many of the specific patterns of differences cross fiction/nonfiction lines. Overall, pronoun use is overwhelmingly more female than male in both fiction and nonfiction. While there are some exceptions with regard to individual pronouns, which will be discussed below, this pattern holds overall for each of first-person, second-person and third-person pronouns in both fiction and nonfiction.

It is evident, however, that it is primarily forms of the pronouns *I*, *you* and *she*, which are in fact used significantly more by females. (It should be noted that the possessive and reflexive forms obey the same distribution as the respective underlying base forms.) Of these, the difference between male and female use of second-person pronouns in both fiction and nonfiction is the most striking and perhaps surprising. The histogram shown in Figure 1 illustrates this point in striking fashion. Note that of the 146 documents in which *you* appears with frequency less than 125, two thirds are male-authored, while of the 110 documents in which *you* appears with frequency greater than 125, two thirds are female-authored.

In functional terms, the use of the second-person pronoun suggests, of course, the drawing of the reader into the text. Similarly, the significant difference between males and females in usage of singular first-person pronouns in nonfiction suggests the introduction of the writer into the text. The difference in usage of singular first-person pronouns is somewhat mitigated in fiction, presumably partially neutralized by conventions of narration and dialogue. That is, both men and women writers provide dialogue in fiction, and thereby tend to use first-person pronouns at about the same rate. Especially interesting is the fact that in fiction it is males who use plural first-person pronouns with significantly greater frequency. We will speculate on the reason for this below.

Table 3. *Statistics for different pronoun classes in nonfiction texts*

Feature	Definition	Genre	Female μ ± stderr	Male μ ± stderr	t-test	Female median	Male median	Mann-Whitney U test
First person	*I, me, my, mine, myself, we, us, our, ours, ourselves*	Nonfiction	149 ± 14	86 ± 8	p<0.0002	66.7	50.2	p<0.1
First person singular	*I, me, my, mine, myself*	Nonfiction	98.8 ± 11	45.0 ± 6.3	p<0.0001	31.0	18.8	p<0.005
First person plural	*we, us, our, ours, ourselves*	Nonfiction	49.7 ± 4.5	40.9 ± 3.4	n/s	27.8	23.7	n/s
Second person	*you, your, yours, yourself*	Nonfiction	63.9 ± 8.0	30.0 ± 5.2	p<0.0005	16.7	3.9	p<0.0001
Third person	*he, him, his, himself, she, her, hers, herself, they, them, their, theirs, themselves*	Nonfiction	243 ± 11	196 ± 9.7	p<0.0001	209	160	p<0.0001
Third person singular	*he, him, his, himself, she, her, hers, herself*	Nonfiction	145 ± 9.9	114 ± 9.1	n/s	90.2	78.1	n/s
Third person: male	*he, him, his, himself*	Nonfiction	91.1 ± 7.7	95.7 ± 7.5	n/s	54.1	64.3	n/s
Third person: female	*she, her, hers, herself*	Nonfiction	53.8 ± 5.1	18.5 ± 3.5	p<0.0001	29.8	5.60	p<0.0001
It	*it*	Nonfiction	89.1 ± 2.8	86.7 ± 2.4	n/s	85.3	82.9	n/s
Its	*its*	Nonfiction	15.3 ± 0.93	19.0 ± 0.79	p<0.005	12.2	19.0	p<0.0001
Third person plural	*they, them, their, theirs, themselves*	Nonfiction	97.8 ± 4.6	81.8 ± 2.7	p<0.005	83.9	78.8	p<0.05

Table 4. *Statistics for different pronoun classes in fiction texts*

Feature	Definition	Genre	Female μ ± stderr	Male μ ± stderr	t-test	Female median	Male median	Mann-Whitney U test
First person	*I, me, my, mine, myself, we, us, our, ours, ourselves*	Fiction	289 ± 12	286 ± 16	n/s	257	218	$p<0.05$
First person singular	*I, me, my, mine, myself*	Fiction	246 ± 10	230 ± 15	n/s	224	180	$p<0.001$
First person plural	*we, us, our, ours, ourselves*	Fiction	42.9 ± 3.2	56.3 ± 3.5	$p<0.01$	33.8	45.8	$p<0.0002$
Second person	*you, your, yours, yourself*	Fiction	161 ± 5.2	119 ± 4.5	$p<0.0001$	161	115	$p<0.0001$
Third person	*he, him, his, himself, she, her, hers, herself, they, them, their, theirs, themselves*	Fiction	683 ± 19	559 ± 15	$p<0.0001$	712	574	$p<0.0001$
Third person singular	*he, him, his, himself, she, her, hers, herself*	Fiction	606 ± 20	459 ± 15	$p<0.0001$	632	469	$p<0.0001$
Third person: male	*he, him, his, himself*	Fiction	271 ± 9.3	305 ± 11	$p<0.05$	276	305	$p<0.05$
Third person: female	*she, her, hers, herself*	Fiction	334 ± 17	154 ± 10	$p<0.0001$	392	128	$p<0.0001$
It	*it*	Fiction	124 ± 2.3	128 ± 2.9	n/s	124	130	n/s
Its	*its*	Fiction	6.87 ± 0.57	10.4 ± 0.89	$p<0.005$	5.3	7.9	$p<0.0005$
Third person plural	*they, them, their, theirs, themselves*	Fiction	77.6 ± 3.2	100 ± 3.8	$p<0.0001$	67.8	92.1	$p<0.0001$

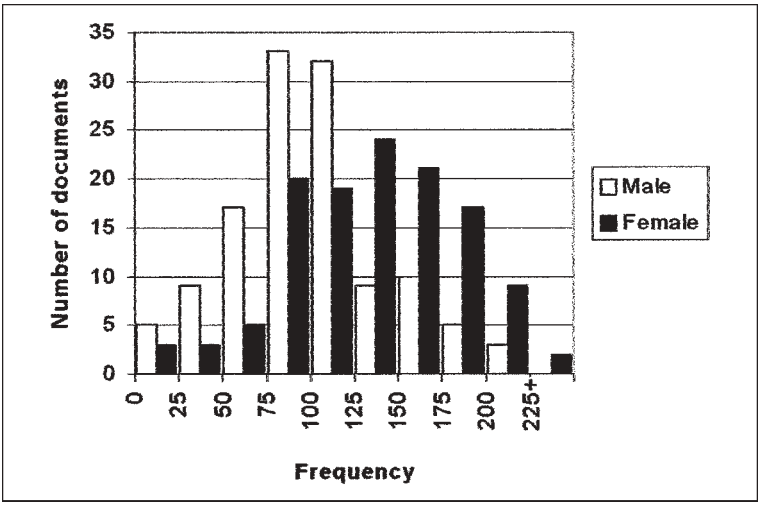

Figure 1. *Per-document frequency of use of the word* you *by male and female authors in fiction documents*[4]

In the case of third-person pronouns, it should be noted that the sum of pronouns generally marked for gender, that is, personal, third-person pronouns (*he, she*) is far greater for females than males in both fiction and nonfiction (there is a particularly striking difference for the female pronouns). By contrast, *it*, which is never personal, is used in equal amounts by males and females and *its* is used more by males in both fiction and nonfiction. This is perhaps to be expected since *its* is both impersonal (as opposed to *his* and *her*) and is a type of specifier.

While the overall pattern of greater usage of pronouns by female authors is clear, there are two types of exceptions that bear closer scrutiny: male authors use more plural pronouns (*we, us, they, them*) in fiction and more male third-person pronouns (*he, him*) in both fiction and nonfiction. With regard to plural pronouns in fiction, we find a consistent pattern across first-, third- and even second-person pronouns. For first-person pronouns, the mean proportion of plural pronouns to overall pronouns (1p – plu / 1p) for male authors is 50.7, while for female authors it is only 42.2. Likewise for third-person pronouns, the mean proportion of plural pronouns to overall pronouns for male authors is 20.4, while for female authors it is only 14.8. For second-person pronouns, where the morphological neutralization of the singular-plural distinction prevents an analogous computation, we used the proportion *yourselves* / (*yourselves* + *yourself*) as a proxy. For males the mean is 6.8, while for females it is only 4.7, which is consistent with the pattern of males using a higher proportion of plurals. Moreover, although the BNC tag system does not distinguish between animate *they* and

inanimate *they*, a hand-count of over 1000 randomly-selected appearances of *they* reveals that the differences in usage of *they* between male and female authors are significant specifically with regard to animate *they*. Thus we may speculate that the greater use of plural pronouns reflects the tendency of male authors to encode classes rather than individualized entities and may also serve as a depersonalization mechanism that reduces the specificity of reference to gender, number, and personhood.

With regard to male third-person pronouns, a hand count of 1000 unique proper nouns reveals that this is due to more references by male authors to male characters in both fiction and nonfiction. One hypothesis that can be ruled out is that in nonfiction *he* is more likely to be used by male authors than by female authors as the unmarked or default third-person pronoun. This turns out not to be the case in our corpus. Specifically, a hand-count of 1000 randomly chosen appearances of *he* reveals that among male authors approximately 10.4% of appearances of *he* are generic, while among female authors 17.0% are generic. Moreover, while the mean frequency of the phrase *he or she* is 1.5 times greater for female authors than for men, the total number of such usages is small (less than 2% of overall usage of *he*) and does not significantly impact the overall numbers. We did not analyze this phenomenon chronologically but it is likely that as the number of 'reformed' female authors (Khosrohashi 1989) increases, the use of generic *he* among female authors will decrease.

In summary, we find here two related aspects of language use that distinguish texts written by females from those written by males. First, female writers use more pronouns that encode the relationship between the writer and the reader (especially first person singular and second person pronouns), while males tend to not to refer to it. Second, female writers more often use personal pronouns that make explicit the gender of the 'thing' being mentioned (third person singular personal pronouns), while males have a tendency to prefer more generic pronouns. Both of these aspects might be seen as pointing to a greater 'personalization' of the text by female authors.

Similar linguistic phenomena have been noted in previous work on male and female linguistic markers. Gender-based variation of the first-person pronoun *I* (and related phrases such as *I think*) has been studied in speech (Holmes 1990; Preisler 1986; Rayson, et al. 1997) and in correspondence (Palander-Collin 1999) and has proven to be a stable difference between male and female language in speech and correspondence; our results extend this to the realm of formal written texts. In particular, Palander-Collin (1999) studied the phrase *I think* and similar evidential phrases in seventeenth century correspondence, and found that in women's letters '[t]he writer and the addressee are both overtly included in the communication

situation and the writer's personal attitude is frequently expressed', a conclusion which accords with our finding in formal written texts that female authors include both the writer and the reader explicitly in the text (even though, unlike in correspondence, the reader is not specifically known). More broadly, as mentioned above, Holmes (1993) has proposed as a possible sociolinguistic 'universal' that females tend to use linguistic devices that stress solidarity between the speaker and listener (Holmes 1984; Holmes 1988; Tannen 1990). To accomplish this, however, it is necessary, especially in formal written texts, to encode the speaker/writer and the listener/reader specifically into the discourse. It is precisely such an encoding that we have found for female authors, with male authors tending to use strategies, which reduce or eliminate such encoding.

6. 'Involvedness' in female writing

Palander-Collin (1999) analyzed her results within the framework devised by Biber (1995), who identified a number of stylistic dimensions based on a multivariate analysis of a set of 67 predetermined linguistic variables. In particular, Palander-Collin found strong evidence for gender-based variation along Biber's (1995) 'dimension 1', finding that women's letters tend to have a more 'involved' style than men's. (As we have noted, it is notoriously difficult to unambiguously map given linguistic markers to communicative function; we use the terms 'involved' and 'informational' as does Biber —simply as a suggestive label for a correlated set of lexical features.) 'Involved' documents contain features, which typically show interaction between the speaker/writer and the listener/reader, such as first and second person pronouns for which we found significant gender differences. Indeed, Biber et al. (1998) also found strong and consistent differences between male and female authors along their 'dimension 1' in English correspondence, with female authors tending to the 'involved' and male authors to the 'informational'. In addition, prominent characteristics of 'involved' writing, other than pronouns, listed in that work are analytic negation, contractions, and present-tense verbs. In Table 5, we show the frequencies of each of these features in our corpus for male and female writing. As is evident, the indicators of 'involvedness' appear with significantly greater frequency in female writing. Note however that the greater use of present-tense verbs by females is neutralized in fiction. Our results are thus consistent with earlier results regarding the 'involvedness' of female-authored texts, but we have also found evidence for specific strategies used by male authors, which seek to reduce the "involvedness" of the text.

Table 5. *Statistics for other 'involved' features in fiction and nonfiction texts*

Feature	Definition	Genre	Female $\mu \pm$ stderr	Male $\mu \pm$ stderr	t-test	Female median	Male median	Mann-Whitney U test
Negative particle	XX0	Nonfiction	63.3 ± 2.5	56.3 ± 1.8	$p<0.05$	57.6	52.0	$p<0.05$
Contractions	Words ending in: n't, 'll, 'd, 're, 've	Nonfiction	26.7 ± 3.4	10.7 ± 1.6	$p<0.0001$	6.60	3.30	$p<0.0001$
Present tense verbs	VVB, VVG, VVZ	Nonfiction	303 ± 9.9	259 ± 7.8	$p<0.001$	299	252	$p<0.005$
Negative particle	XX0	Fiction	123 ± 2.7	104 ± 3.1	$p<0.0001$	125	99.4	$p<0.0001$
Contractions	Words ending in: n't, 'll, 'd, 're, 've	Fiction	153 ± 5.7	126 ± 5.4	$p<0.001$	162	123	$p<0.0005$
Present tense verbs	VVB, VVG, VVZ	Fiction	315 ± 7.3	322 ± 11	n/s	306	289	n/s

7. Male markers: Specifiers

Male authors also have clear distinguishing markers. The more frequent use of determiners by male authors is not, as might be suspected, merely a consequence of their (slightly) greater use of common nouns. In fact, the difference in mean value of the proportion determiners/common nouns is significant both for fiction and for nonfiction (see Table 6). This suggests that male authors are more likely to 'indicate' or 'specify' the things that they write about. Indeed, the greater use of determiners in male writing is not an isolated phenomenon. Similar differences in use are obtained for other language forms, which serve to specify which particular 'things' in the world (as encoded in nouns) are being written about. We find that male authors reliably provide more specification. Although we cannot explore the issue by automatic means, examination of the texts suggests that the use of determiners reflects that male writers are mentioning classes of things in contrast to female writers who are personalizing their messages and use pronouns to link one mention of a person or object to other mentions.

Table 6 shows results for a variety of specification features, which were suggested by features found by our automatic learning procedure. In both fiction and nonfiction, we find male authors using more post-head noun modification with an *of*-phrase (e.g., 'garden of roses'). In fiction, male authors quantify things more often by using cardinal numbers in a noun phrase. This phenomenon is neutralized in nonfiction possibly due to the greater quantification inherent to most nonfiction genres. Similarly, the greater use of attributive adjectives by male authors in nonfiction writing is attenuated in fiction writing, likely due to conventions of the genre. Finally, as noted earlier, the pronoun *its*, which serves to specify the identity or properties of a thing, occurs with far greater frequency in male-authored texts, both fiction and nonfiction.

In terms of Biber's dimensions, specifier use relates primarily to the 'informational' half of his 'dimension 1'. Our results thus confirm and extend his and others' findings (Mulac and Lundell 1994; Biber et al. 1998) that male authors tend to use more 'informational' features. In particular, prepositions are among the features considered to be 'informational'. We found an especially strong difference in one case where a prepositional phrase conclusively functions as a noun modifier (noun followed by *of*). Attributive adjectives are found by Biber to be both 'informational' and 'non-narrative' ('dimension 2'), which indicates that male writing and nonfiction may share both such features. Quantification (reasonably considered an 'informational' feature) is not considered by Biber; however, our results here support the related observation (Mulac et al. 1990; Mulac and Lundell 1994) that 'references to quantity or place' is a male indicator in short student essays.

Table 6. *Statistics for nominal specifiers in fiction and nonfiction texts*

Feature	Definition	Genre	Female μ ± stderr	Male μ ± stderr	t-test	Female median	Male median	Mann-Whitney U test
Determiners	AT0, DT0	Nonfiction	1152 ± 12	1247 ± 9.0	p<0.0001	1149	1247	p<0.0001
Determiner/Noun	100*Det / NN	Nonfiction	57.6 ± 0.59	61.1 ± 0.55	p<0.0001	58.0	61.0	p<0.0001
Cardinal numbers	CRD_NN, CRD_AJ0, CRD_PRF	Nonfiction	57.0 ± 2.0	60.3 ± 2.3	n/s	50.5	54.6	n/s
Attribute Attributive adjectives	ADJ_NN, ADJ_ADJ	Nonfiction	451 ± 10	514 ± 9.8	p<0.0001	438	505	p<0.0001
Noun-of	NN_PRF	Nonfiction	278 ± 8.1	327 ± 6.6	p<0.0001	269	328	p<0.0001
Determiners	AT0, DT0	Fiction	908 ± 13	1041 ± 10	p<0.0001	889	1047	p<0.0001
Determiner/Noun	100*Det / NN	Fiction	61.7 ± 0.48	65.7 ± 0.48	p<0.0001	61.5	65.7	p<0.0001
Cardinal numbers	CRD_NN, CRD_AJ0, CRD_PRF	Fiction	35.7 ± 1.4	48.8 ± 2.0	p<0.0001	31.3	43.7	p<0.0001
Attribute Attributive adjectives	ADJ_NN, ADJ_ADJ	Fiction	267 ± 5.5	280 ± 7.7	n/s	256	273	n/s
Noun-of	NN_PRF	Fiction	134 ± 4.1	148 ± 4.5	p<0.05	130	151	p<0.005

Similarly, Johnstone (1993) observed that in oral narratives, male narrators gave more references to place and time than female narrators. Prominent characteristics of informational writing listed in 'dimension 1' that are not directly linked to specification are word length and type/token ratio. Results for these features on our corpus are shown in Table 7. These results are consistent with the hypothesis that male writing tends to exhibit more 'informational' features. Note that, possibly due to conventions of the nonfiction genres, the higher type/token ratio found in male fiction is neutralized in nonfiction.

We did not find evidence of specific strategies used by female authors to reduce specification analogous to the evidence found for male strategies reducing personalization discussed above. However, it may be that the generally higher use by females of pronouns serves to maintain a higher degree of continuity among the 'things' in a text, and so reduces the need to use specification (compare recent work by Cheshire [2002]).

8. Gender and genre

Our results about pronouns and determiners may be generalized in yet another direction. Although the nonfiction documents in our corpus come from a variety of widely-differing genres, certain significant statistical differences between the fiction and nonfiction documents in the corpus are clear. As a glance at Table 2 indicates, pronouns appear with overwhelmingly greater frequency in fiction (928 per 10,000 words) than in nonfiction (336 per 10,000 words). Conversely, determiners appear with much greater frequency in nonfiction (1200 per 10,000 words) than in fiction (974 per 10,000 words). This immediately suggests a correlation between female-male and fiction-nonfiction differences. We examined this hypothesis by considering all the features used in our experiments (limiting ourselves to the most frequent for reliability). In Figures 2 and 3, we plot—for each of the 100 most frequent function words and the 100 most frequent part-of-speech (POS) n-grams, respectively—the surplus of the feature in male writing (X-axis) against the surplus of the feature in nonfiction (Y-axis). As is evident from the almost linear flow of the plot, the correlation of male (female) writing characteristics with characteristics of nonfiction (fiction) goes well beyond the bounds of the features we have examined above. Pearson's correlations are shown in Table 8, demonstrating conclusively that a strong relationship exists.

It should be noted, though, that in the case of POS, the plotted points (features) are not independent of each other since the same parts of speech may be used in a number of n-grams. In fact, all the features in the extreme

Table 7. *Statistics for other 'informational' features in fiction and nonfiction texts*

Feature	Definition	Genre	Female μ ± stderr	Male μ ± stderr	t-test	Female median	Male median	Mann-Whitney U test
Nouns	NN, NP0	Nonfiction	2235 ± 25	2293 ± 18	n/s	2248	2321	n/s
Prepositions	PRP, PRF	Nonfiction	1143 ± 15	1211 ± 10	$p<0.0005$	1148	1226	$p<0.0005$
Word length		Nonfiction	4.64 ± 0.023	4.79 ± 0.020	$p<0.0001$	4.65	4.81	$p<0.0001$
100 * type / token[a]		Nonfiction	15.8 ± 0.57	14.7 ± 0.52	n/s	13.2	12.8	n/s
Nouns	NN, NP0	Fiction	1677 ± 23	1822 ± 19	$p<0.0001$	1638	1801	$p<0.0001$
Prepositions	PRP, PRF	Fiction	829 ± 11	867 ± 11	$p<0.05$	809	868	$p<0.005$
Word length		Fiction	4.13 ± 0.012	4.16 ± 0.017	n/s	4.12	4.18	$p<0.01$
100 * type / token[a]		Fiction	12.0 ± 0.49	13.6 ± 0.55	$p<0.05$	10.6	12.1	$p<0.0001$

[a]As per Biber (1995) we counted the number of different words in the first 400 words of each document, and then divided by 4. This balances the fact that longer documents are likely to have fewer word types per word

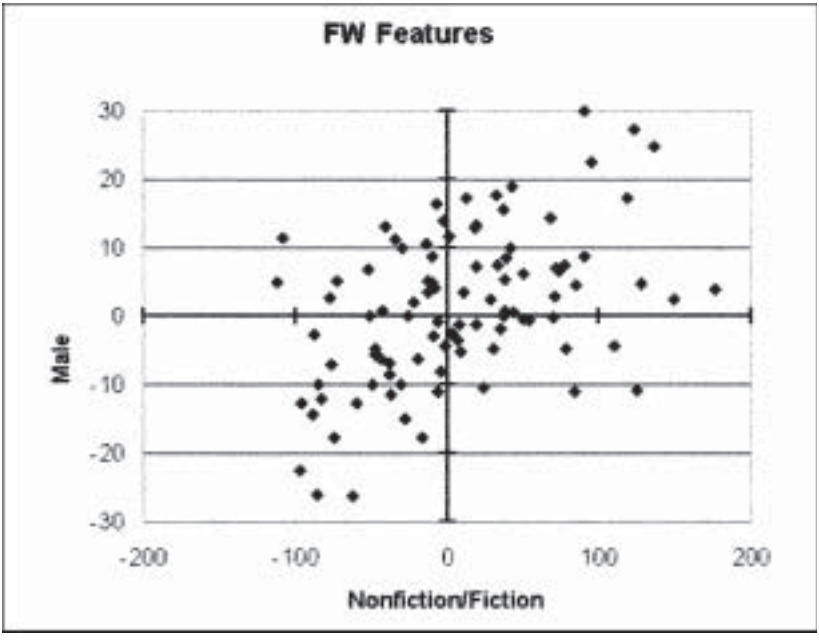

Figure 2. *Scatterplot showing normalized frequency differences (gender versus genre) for the most frequent 100 function word (FW) features*

upper right (male/nonfiction) corner of each graph were related to prepositions and determiners, and all the features (with a single exception) in the extreme lower left (female/fiction) corner of each graph were related to pronouns. The single example of a non-pronoun feature, which is both overwhelmingly prevalent in fiction and in female writing is: PUN_PUQ—punctuation followed by quotation marks (typical of end quotes). This suggests that the use of dialogue, typical of fiction, may also a characteristic of female writing. Alternatively, the use of quotation marks after punctuation, particularly in nonfiction, indicates that the female texts introduce other people's words into their writing more than the male texts do, as has already been observed with regard to oral narration (Johnstone 1993).

9. Sample texts

Let us now consider several illustrative passages. First, we consider opening passages of two articles published in the same journal (*Language and Literature*), one written by a male author (Paul Simpson) and one by a female author (Diane Blakemore).

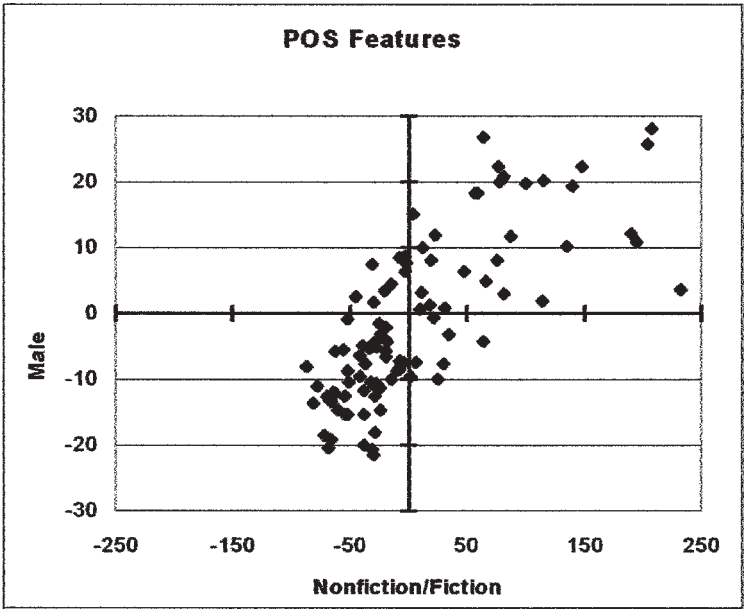

Figure 3. *Scatterplot showing normalized frequency differences (gender versus genre) for the most frequent 100 part-of-speech (POS) features*

Table 8. *Pearson's correlation between normalized genre and sex differences for 100 most frequent FW and POS features, respectively*

Feature set	Correlation	95% Confidence interval	Significance
FW	0.56	0.36, 0.71	$p<0.0001$
POS	0.76	0.62, 0.85	$p<0.0001$

(1) (*Language and Literature*, vol. 1 [1992]. Paul Simpson)
The main aim of this article is to propose an exercise in stylistic analysis which can be employed in the teaching of English language. It details the design and results of a workshop activity on narrative carried out with undergraduates in a university department of English. The methods proposed are intended to enable students to obtain insights into aspects of cohesion and narrative structure: insights, it is suggested, which are not as readily obtainable through more traditional techniques of stylistic analysis. The text chosen for analysis is a short story by Ernest Hemingway comprising only 11 sentences. A jumbled version of this story is presented to students who are asked to assemble a cohesive and well formed version of the story. Their

re-constructions are then compared with the original Hemingway version.

(2) (*Language and Literature*, vol. 2 [1993]. Diane Blakemore)
My aim in this article is to show that given a relevance theoretic approach to utterance interpretation, it is possible to develop a better understanding of what some of these so-called apposition markers indicate. It will be argued that the decision to put something in other words is essentially a decision about style, a point which is, perhaps, anticipated by Burton-Roberts when he describes loose apposition as a rhetorical device. However, he does not justify this suggestion by giving the criteria for classifying a mode of expression as a rhetorical device. Nor does he specify what kind of effects might be achieved by a reformulation or explain how it achieves those effects. In this paper I follow Sperber and Wilson's (1986) suggestion that rhetorical devices like metaphor, irony and repetition are particular means of achieving relevance. As I have suggested, the corrections that are made in unplanned discourse are also made in the pursuit of optimal relevance. However, these are made because the speaker recognises that the original formulation did not achieve optimal relevance. In contrast, deliberate reformulations are designed to achieve particular contextual effects, and they should not be taken to indicate a failure to communicate any more than, for example, repetition.

Already from the first phrase of each passage, we might venture a guess, which is which. Indeed, it is the female Blakemore who writes 'My aim', while the male Simpson uses the less personal and more specified 'The main aim'. Blakemore further personalizes by using the phrases 'I follow' and 'As I have suggested'. Simpson, by contrast, uses only a single personal pronoun in the whole passage and it is plural. Moreover, after introducing Burton-Roberts, Blakemore emphasizes his personhood by following up twice with references to 'he'. By contrast, Simpson, having referred to Hemingway, makes no effort to personalize and refers subsequently only to 'Hemingway's version'. In addition, Blakemore's use of twelve present tense active verbs (base form, _s), as opposed to Simpson's use of only three, effectively places the actors at the center of her narrative.

Furthermore, in six sentences Simpson uses eight *of* phrases to modify nouns (e.g., 'more traditional techniques of stylistic analysis'), while in eight sentences, Blakemore uses only six *of* modifiers. Finally, Blakemore uses four negatives (*not, nor*), while Simpson uses only one. It appears that wording propositions in the negative is another device for relating to the reader by setting up a contrast with the reader's expected state of the world (e.g., 'they should not be taken to indicate a failure to communicate . . .').

Let us now consider two fiction passages. The following passages are the respective opening passages of two novels (*Saigon* by Anthony Grey and *Jerusalem the Golden* by Margaret Drabble) each centered on the protagonist's move to a new city, Saigon and London, respectively.

(3) (*Saigon*, Anthony Grey)
By 1925 present-day Vietnam was divided into three parts under French colonial rule. The southern region embracing Saigon and the Mekong delta was the colony of Cochin-China; the central area with its imperial capital at Hue was the protectorate of Annam; and the northern region, Tongking, was also a separate protectorate with its capital at Hanoi. The Annamese emperor, Khai Dinh, in theory ruled the two northern regions from Hue with the benefit of French protection, while Cochin-China was governed directly from Paris but in effect all three territories were ruled as colonies. Some backward tribes inhabited the remoter mountains and jungles but the main population was of the same race; today they are known as Vietnamese but then the outside world knew them as Annamites or Annamese. They had detached themselves from the torrent of peoples that in prehistory had poured out of China onto the countless islands of the Pacific and, settling the eastern coastal strip of the Indochina.

(4) (*Jerusalem the Golden*, Margaret Drabble)
Clara never failed to be astonished by the extraordinary felicity of her own name. She found it hard to trust herself to the mercy of fate, which had managed over the years to convert her greatest shame into one of her greatest assets, and even after years of comparative security she was still prepared for, still half expecting the old gibes to be revived. But whenever she was introduced, nothing greeted the amazing, all-revealing Clara but cries of "How delightful, how charming, how unusual, how fortunate," and she could foresee a time when friends would name their babies after her and refer back to her with pride as the original from which inspiration had first been drawn. Finally her confidence grew to such an extent that she was able to explain that she had been christened not in the vanguard but in the extreme rearguard of fashion, after a Wesleyan great-aunt, and that her mother had formed the notion not as an unusual and charming conceit but as a preconceived penance for her daughter, whose only offences at that tender age were her existence and her sex.

These passages illustrate in extreme fashion the fundamental differences borne out by our statistical findings. Grey opens his book with a recitation of facts; Drabble opens hers with her protagonist's thoughts. Consequently,

Drabble uses 17 singular feminine pronouns, while Grey uses only four animate pronouns altogether and all are plural. In his 161 words, Grey uses 46 proper or common nouns, while Drabble uses only 33 in 187 words. Grey uses 4 numbers, Drabble none. Grey uses the determiner *the* 18 times, Drabble only 9 times. Overall, one could easily imagine Grey's introductory passage in a nonfiction work, while Drabble's passage is unmistakably fiction.

10. Conclusions

The results presented above offer convincing evidence that there are indeed different strategies employed by men and women in setting forth information and especially in encoding the relation between writer and reader in texts. Ascertaining the precise communicative functions and broader social significance of these respective linguistic strategies is a difficult and ideologically-loaded problem which is beyond the scope of this article. Nevertheless, the fact that these results extend findings substantiated independently in less formal communication contexts to large formal written texts intended for an unseen audience over a range of genres is very suggestive. The extension to low-interaction linguistic modalities invites a re-examination of the mechanisms of socialization of men and women into interactional styles and related differences in the use of language and hints at the possibility that new learning and other cognitive explanations may be called for. For example, recent physiological studies (see Canli et al. 2002) point to a difference in men's and women's processing of emotional material that may be indirectly related to the findings in the use of language.

In addition to socialization into gender, there is also an important gender-genre issue to be explored. The strong correlation between male/female differences and nonfiction/fiction differences suggests that different writers involve themselves and the information they are presenting into the different social processes found in the culture. The distribution of the encodings of different meanings cuts across both gender and genre in clear ways that requires more consideration of register issues.

The consistent differences over millions of words suggest the large amount of work still necessary to understand how different writers develop a style reflected by a series of linguistic features that is then parallel to the genre differences that are recognized and recognizable in a speech community. Do males and females read different kinds and amounts of text? Are they invited to imitate some texts rather than other texts? Do the meanings in some texts, as encoded by the particular sets of linguistic features, resonate with different views of the world? These are just some of the questions

that need careful exploration through the detailed analysis of the specific linguistic characteristics of texts.

The process through which different writing styles develop and how they relate to their social context remains a topic for much further research—but the existence of such differences would appear now to be firmly established. It remains for further study to determine the extent to which these distinctions remain consistent across cultural and chronological lines.

Appendix : Journal and fiction passages

Blakemore, Diane (1993). The relevance of reformulations. *Language and Literature* 2 (2): 101–119.
Drabble, Margaret (1988). *Jerusalem the Golden*. London: Penguin Group.
Grey, Anthony (1983). *Saigon*. London: Pan Books.
Simpson, Paul (1992). Teaching stylistics: Analysing cohesion and narrative structure in a short story by Ernest Hemingway. *Language and Literature* 1 (1): 47–65.

Notes

1. A complete listing of the documents used in this study may be accessed via the web page at <http://www.ir.iit.edu/~argamon/gender.html>.
2. (A full listing of the features used in this study can be found on the web site <http://www.ir.iit.edu/~argamon/gender.html>.
3. Significance of the differences was tested both using students' t-test for independent samples (with Welch's approximation for unequal variances) as well as the non-parametric Mann-Whitney U test. All feature frequencies in this article are given per 10,000 equivalent tokens (words or part-of-speech n-grams).
4. In the histogram, the height of the vertical bars indicates the number of documents with frequency of *you* in the indicated range.

References

Aries, E. J. and Johnson, F. L. (1983). Close friendship in adulthood: Conversational content between same-sex friends. *Sex Roles*, 9 (12): 1183–1196.
Baayen, H., van Halteren, H, and Tweedie, F. (1996). Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing*, 11.
Berryman-Fink, C. L. and Wilcox, T. R. (1983). A multivariate investigation of perceptual attributions concerning gender appropriateness in language. *Sex Roles* 9: 663–681.
Bergvall, V., Bing, J., and Freed, A. (eds.) (1996). *Rethinking Language and Gender Research: Theory and Practice*, New York: Addison Wesley Longman.

Biber, D. (1988). *Variation Across Speech and Writing*. Cambridge: Cambridge University Press.

—(1995). *Dimensions of Register Variation: A Cross-linguistic Comparison*. Cambridge: Cambridge University Press.

Biber, D., Conrad S., and Reppen, R. (1998). *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Bing, J. and Bergvall, V. (199F6) The question of questions: beyond binary thinking. In *Rethinking Language and Gender Research: Theory and Practice* V. Bergvall, J. Bing, and A. Freed. (eds.), New York: Addison Wesley Longman.

Cameron, D., McAlinden, F., and O'Leary, K. (1988). Lakoff in context: the social and linguistic function of tag questions. In *Women in their Speech Communities*, J. Coates and D. Cameron (eds.) 74–93. New York: Longman.

Canli, T., Desmond, J.E., Zhao, Z., and Gabrieli, D. E. (2002). Sex differences in the neural basis of emotional memories. *Proceedings of the National Academy of Science* 99: 10789–10794.

Chambers, J. C. (1992). Linguistic correlates of gender and sex. *English World-Wide,* 13 (2): 173–218.

Cheshire, J. (2002). Information structure in male and female adolescent talk. *Journal of English Linguistics* 30 (2): 217–238.

Coates, J. (ed.) (1998). *Language and Gender: A Reader*. Oxford: Blackwell.

Dagan,I., Karov, Y., and Roth, D. (1997). Mistake-driven learning in text categorization, in *EMNLP-97: 2nd Conference on Empirical Methods in Natural Language Processing* 55–63.

Eckert, P. (1997). Gender and sociolinguistic variation. In *Readings in Language and Gender*. J. Coates (ed.), 64–75. Oxford: Blackwell.

Halliday, M. A. K. (1994). *An Introduction to Functional Grammar*. 2nd edition. London: Edward Arnold.

Herring, S. (1996). Two variants of an electronic message schema: In *Computer-Mediated Communication: Linguistic, Social and Cross-Cultural Perspectives*, S. Herring (ed.), 81–106. Amsterdam /Philadelphia: John Benjamins.

Holmes, D. (1998). The evolution of stylometry in humanities scholarship. *Literary and Linguistic Computing*, 13 (3): 111–117.

Holmes, J. (1984). Women's language': A functional approach. *General Linguistics* 24 (3): 149–178.

—(1988). Paying compliments: A sex-preferential positive politeness strategy. *Journal of Pragmatics*. 12 (3): 445–465.

—(1989). Sex differences and apologies: One aspect of communicative competence. *Applied Linguistics*. 10 (2): 194–213.

—(1990). Hedges and boosters in women's and men's speech. *Language and Communication* 10 (3).

—(1993). Women's talk: The question of sociolinguistic universals. *Australian Journal of Communications* 20 (3): 125–149.

Johnstone, B. (1993). Community and contest: Midwestern men and women constructing their worlds in conversational storytelling. In *Gender and Conversational Interaction*, D. Tannen (ed.), 62–80. Oxford: Oxford University Press.

Key, M. R. (1975). *Male/Female Language*. Metuchen, NJ: Scarecrow Press.

Khosroshahi (1989), Penguins don't care, but women do: A social identity analysis of a Whorfian problem. *Language in Society* 18 (4): 505–525.

Kivinen, J. and Warmuth, M. (1997). Additive versus exponentiated gradient updates for linear prediction. *Information and Computation* 132 (1): 1–64.

Koppel, M., Argamon, S., and Shimoni A. R. (2001). Automatically determining the gender of a text's author. Bar-Ilan University Technical Report BIU-TR-01-32.
Labov, W. (1990). The intersection of sex and social class in the course of linguistic change. *Language Variation and Change* 2: 205–254.
Lakoff, R. T. (1975). *Language and Women's Place*. New York: Harper Colophon Books.
Lewis, D., Schapire, R., Callan, J., and Papka, R. (1996). Training algorithms for text classifiers. *Proceedings of the 19th ACM/SIGIR Conference on Research and Development in IR* 306–298.
Littlestone, N. (1987). Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning* 2 (4): 285–318.
McEnery, T. and Wilson, A. (1996). *Corpus Linguistics*. Edinburgh: Edinburgh University Press.
Mulac, A. and Lundell, T. L. (1994). Effects of gender-linked language differences in adults' written discourse: Multivariate tests of language effects. *Language and Communication* 14 (3): 299–309.
Mulac, A., Studley, L. B., and Blau, S. (1990). The gender-linked language effect in primary and secondary students' impromptu essays. *Sex Roles* 23 (9/10): 439–469.
Palander-Collin, M. (1999). Male and female styles in 17th century correspondence. *Language Variation and Change* 11: 123–141.
Preisler, B. (1986). *Linguistic Sex Roles in Conversation.* Berlin/New York: Mouton de Gruyter.
Rayson, P., Leech, G., and Hodges, M. (1997). Social differentiation in the use of English vocabulary: Some analyses of the conversational component of the British National Corpus. *International Journal of Corpus Linguistics* 2: 133–152.
Sebastiani, F. (to appear). Machine learning in automated text categorization. *ACM Computing Surveys*.
Simkins-Bullock, J. A. and Wildman, B. G. (1991). An investigation into the relationship between gender and language. *Sex Roles* 24: 149–160.
Stamatatos, E., Fakotakis, N., and Kokkinakis,G . (2001). Computer-based authorship attribution without lexical measures. *Computers and the Humanities* 35: 193–214.
Tannen, D. (1990). Gender differences in topical coherence: Creating involvement in best friends' talk. *Discourse Processes* 13: 73–90.
Trudgill, P. (1972). Sex, covert prestige and linguistic change in the urban British English of Norwich, *Language in Society* 1: 179–195.

Shlomo Argamon is Associate Professor of Computer Science at the Illinois Institute of Technology. He received his Ph.D. in computer science from Yale University in 1994 and was a Fulbright Scholar at Bar-Ilan University from 1994 to 1996. His current research includes machine learning, computational stylistics, and language learning. Address for correspondence: Department of Computer Science, Illinois Institute of Technology, 10 w. 31st St., Chicago, IL 60616, USA, <argamon@iit.edu>.

Moshe Koppel is Associate Professor of Computer Science at Bar-Ilan University in Ramat-Gan, Israel. He received a Ph.D. in mathematics from Courant Institute in New York in 1979 and was a member of the Institute for Advanced Study in Princeton in 1980. His research interests include machine learning, text categorization and Rabbinic logic. Address for correspondence: Department of Mathematics and Computer Science, Bar-Ilan University, Ramat Gan 52900, Israel, <koppel@netvison.net.il>.

Jonathan Fine (PhD linguistics, Cornell University) has done research on children's language, genre, the language of psychiatric patients and functional systemic linguistics.

He is associate professor of linguistics in the English department of Bar-Ilan University. His current research includes the language of bilinguals with schizophrenia, bilingual specific language impairment, and the language of autism and attention-deficit hyperactivity disorder. Address for correspondence: Department. of English, Bar-Ilan University, Ramat Gan 52900, Israel, <finejo@mail.biu.ac.il>.

Anat Rachel Shimoni received her M.Sc. in computer science from Bar-Ilan University. Her thesis concerned style-based text categorization. She is now a member of the research staff at NDS in Jerusalem. Address for correspondence: Anat Rachel Shimoni, Department of Mathematics and Computer Science, Bar-Ilan University, Ramat Gan 52900, Israel, <averkar@ndsisrael.com>.