Handbook of Research on Discourse Behavior and Digital Communication:
Language Structures and Social Interaction

Rotimi Taiwo
Obafemi Awolowo University, Nigeria

Volume I

Information Science
REFERENCE

INFORMATION SCIENCE REFERENCE
Hershey · New York

Director of Editorial Content:	Kristin Klinger
Director of Book Publications:	Julia Mosemann
Acquisitions Editor:	Lindsay Johnston
Development Editor:	Christine Bufton
Typesetter:	Gregory Snader
Production Editor:	Jamie Snavely
Cover Design:	Lisa Tosheff
Printed at:	Yurchak Printing Inc.

Published in the United States of America by
Information Science Reference (an imprint of IGI Global)
701 E. Chocolate Avenue
Hershey PA 17033
Tel: 717-533-8845
Fax: 717-533-8661
E-mail: cust@igi-global.com
Web site: http://www.igi-global.com/reference

Library of Congress Cataloging-in-Publication Data

Handbook of research on discourse behavior and digital communication : language structures and social interaction / Rotimi Taiwo, editor. p. cm.

Includes bibliographical references and index.

Summary: "This book provides a variety of chapters on discourse behavior in digital communication, including computer-mediated communication as well as aspects of behavior typically associated with online discourse like cyberbullying"--Provided by publisher. ISBN 978-1-61520-773-2 (hbk.) -- ISBN 978-1-61520-774-9 (ebook) 1. Communication and technology. 2. Digital communications--Social aspects. 3. Telecommunication--Social aspects. 4. Telematics--Social aspects. I. Taiwo, Olurotimi Adebowale. P96.T42H365 2010

302.23'1--dc22

2009036916

British Cataloguing in Publication Data
A Cataloguing in Publication record for this book is available from the British Library.

All work contributed to this book is new, previously-unpublished material. The views expressed in this book are those of the authors, but not necessarily of the publisher.

Chapter 18
First Person Pronouns in Online Diary Writing

John Newman
University of Alberta, Canada

Laura Teddiman
University of Alberta, Canada

ABSTRACT

It is well-known that first person pronouns have a particularly important role to play in conversation. "Online diary" style of writing is less well understood and the role of first person pronouns in that style invites further study. In this chapter the authors explore these pronouns in UK and US online diaries, paying particular attention to frequency and collocational relations. In previous corpus-based studies of English genres, first person pronouns have tended to be considered as one larger set without differentiation. The authors find, on the contrary, that the differences between these forms can be very revealing in the way they distinguish online diary style of writing from other genres such as conversation and fiction writing. The findings underline the need to respect inflectional variants of lemmas as objects of study in their own right.

INTRODUCTION

Online diary writing is one of a number of new genres of written language which have emerged with the increasing accessibility of the Internet. The proliferation of websites which encourage such writing means that data from this genre is relatively easy to obtain for the purposes of academic study. Online diaries are, in fact, written and published on the Internet in the expectation that they will be read by an online audience, as opposed to traditional diary writing which typically is intended to remain private. Presumably, online diary writing tends to have a high degree of author involvement which will translate into relatively high frequency of usage of first person pronouns. Consequently, we focus here on the usage of the first person forms of English pronouns in a corpus of online diary writing. In the approach adopted here, we turn attention away from the lemma (e.g., the category of "first person pronoun") to the individual inflected forms of that category (*I*, *me*, etc.), reflecting a new interest among

DOI: 10.4018/978-1-61520-773-2.ch018

linguists in the differential behaviour of the word forms that make up a lemma.

BACKGROUND

Recent case studies of verbs in English have revealed interesting patterning around particular inflected forms, as opposed to lemmas, suggesting that investigating language at the inflectional level is a promising line of inquiry (cf. Newman & Rice, 2004; Newman, in press).[1]Scheibman (2001), in a study of informal conversation, found that first person singular (1SG) and second person singular (2SG) subjects occur with particular verbs of cognition with a relatively high frequency (*I guess*, *I don't know*, *you know*, *I mean*) reflecting the particular pragmatic role played by these phrases in conversation. Scheibman (2001, p. 84) also emphasizes the need to examine 'local' patterns in grammatical research and cautions against relying just on the superordinate grammatical categories (person, verb type, tense etc.). In a similar way, Tao (2001, 2003) discusses the prominence of the simple present tense forms of the verb REMEMBER, used with a first person singular subject (*I remember*) or a null subject (*remember*), again demonstrating the importance of studying particular inflected forms of a verb, rather than just the lemma. The Scheibman and Tao studies both point to the subject form of the 1SG pronoun in English, *I*, as playing a particularly important role in conversational style.

In light of this previous research, we decided to explore the differential behavior of the first person pronouns in different genres in English. First person pronouns are well known as forms which indicate "an interpersonal focus and a generally involved style" (Biber, 1988, p. 225) and which play an important role in distinguishing spoken and written registers (see, e.g., Biber, 1988, p. 225 for further references). Not surprisingly, Biber (1988) identifies the class of first person pronouns as a "linguistic feature", worthy of inclusion in the 67 features which form the basis for his corpus-linguistic analysis of stylistic variation in English. This class consists of *I*, *me*, *we*, *us*, *my*, *mine, our*, *myself*, *ourselves*, *ours*. Without disputing the value of grouping these forms together as part of Biber's (1988) study, we believe that there is much to be gained, too, from investigating properties of the individual "inflected" forms of this class, in addition to studying these forms collectively, as it were, at the level of the lemma. For the purposes of this study we restrict ourselves to *I*, *me*, *my*, *we*, *us*, and *our*.

We chose to make online diary writing a particular focus of this study. Diary writing, generally, has been a relatively neglected kind of writing in corpus studies. It is not a type of writing that is represented in the British National Corpus (BNC), for example. This is perhaps understandable, since it is writing which prototypically would be for the benefit of the writer alone and so not generally accessible to others. Biographies, personal letters, and email, all of which are represented in the BNC, bear similarities to diary writing, though one would expect a number of differences, too, in style, content, and audience.[2] Online diary writing cannot be equated with traditional diary writing intended only to be read by its author. Nor can it be equated with literary outputs which see publication through established presses. Helen Fielding's 1996 novel *Bridget Jones's Diary*, for example, has a complex origin, arising out of newspaper columns written by Fielding, fashioned into a first-person narrative. Additionally, the author herself points to Jane Austen's *Pride and Prejudice* as a source of inspiration for the book.[3] While *Bridget Jones's Diary* is of interest in its own right, of course, and may even be, in some ways, a model for particular online diarists, its complex origin makes it rather different from typical online diary writing. McNeill (2003, 2005) provides an illuminating review of online diary writing and how it compares with traditional forms of diary writing. She draws attention to the manner in which "the assertion of identity that the online diary performs demands a response, a witness for the confession to be enacted" (McNeill, 2003, p. 36), and how "for online diarists, who write explicitly to be read, the absence of an (active, responsive) audience would be a significant blow" (McNeill, 2003, p. 36). The online diarist, in other words, is not engaged in a purely private, secretive form of writing, but is writing with a view to being read and understood and responded to by an audience, even if it is largely an anonymous one. While feedback to authors can be a significant part of many kinds of internet-based communicative acts (cf. Stefanone & Jang, 2008; Miura & Yamashita, 2007; Lenhart & Fox, 2006), we feel it is particularly relevant in online diary writing, as argued by McNeill above.

THE CURRENT STUDY

Online diary writing, being so accessible, presents a unique opportunity for study. It seemed to us that diary writing would be a particularly interesting genre for the study of first person forms, with perhaps greater first person involvement than in any other genre (cf. the advice offered on style quoted in footnote 2). The online diary corpora used in this study, as well as a number of other studies based on diary writing, have already yielded results on subject ellipsis which are relevant to understanding styles employed in such writing (Teddiman & Newman, 2007; Haegeman & Ihsane, 1999, 2001). We allow ourselves to use the term "genre" to refer to online diary writing, using the term in the relatively fuzzy way advocated by Lee (2001, p. 47), i.e., to "describe groups of texts collected and compiled for corpora or corpus-based studies".

The Online Diary Corpora

The Diary corpora were constructed from publicly available online weblogs (blogs) hosted by a blogging service called Livejournal (http://www.livejournal.com) between July 2006 and December 2006.[4] Users were randomly selected from within two geographical regions, the United States and the United Kingdom. All journals included in the corpus had been recently updated and active at the time of data collection, and included information about the sex and age of the user wherever possible. Fifty online diaries were selected within each region, with approximately 2,000 words collected per user, comparable to the sampling size used for individual texts in the International Corpus of English. The US and UK sub-corpora each contain approximately 100,000 words (US: 102,781, UK: 102,216), with a total of 204,997 words in the whole corpus. Hyper-text mark-up, journal tags, and timestamps were removed from the text and are not included in the word counts.

As noted above, the selection of online diaries was random in this sample. Nevertheless, in both the US and UK populations, more women were identified as diarists than men (65% vs. 35% over both populations). While age varied widely (16-63), the mean age of bloggers was 25 (24.7). See Table 1 for a breakdown of these population demographics.

First Person Forms

As a first step towards understanding the distribution of the singular forms, we summed up the number of *I*, *me*, *my*, *we*, *us*, and *our* tokens which occurred in the diary corpora and compared these results with totals of these forms in four well known and often used genres in corpus linguistics: conversation, fiction, news, and academic writing. For these four genres, we used the four 1 million-word corpora of BNCBaby, representing a sampling of the BNC World Edition, referred to here as Dem ("spoken demographic", i.e., conversations), Fic (fictional prose), News (newspaper texts), and Acad (academic writing from periodicals and books). These are also the genres which figure prominently in Biber et al (2000). In the case of the diary corpora, some pre-editing of the raw

Table 1. Demographic breakdown of the diary corpora by region, gender, and age

	Users (#)	Age		Total Words
		Min – Max	Mean	
US	**50**	**17 – 63**	**24**	**102781**
Male	16	17 – 41	24.1	33135
Female	34	17 – 63	24	69646
UK	**50**	**16 – 52**	**25.4**	**102216**
Male	19	18 - 45	26.7	38942
Female	31	16 – 52	24.7	63274
Total	**100**	**16 – 63**	**24.7**	**204997**
Male	35	17 – 45	25.5	72077
Female	65	16 – 63	24.3	132920

text was done, replacing contracted forms (*I've*, *I'm*, *I'll*, *I'd* etc.) with their full forms to facilitate easier searches. Some occurrences of these contractions were spelled without the apostrophe in the original blogs and these tokens were normalized to ensure that the *I* form was counted correctly in these cases.

Female authors used first person singular pronouns more often than male authors, although the general patterns between the pronouns were the same for both sexes. This result did not extend into the first person plural forms. These results contrast with those of Herring and Paolillo (2006), who found that female authors favoured *we* over male authors, but who did not find any differences in the use of first person singular *I*. These differences could stem from journal selection criteria. In their selection of journals, Herring and Paolillo (2006) preferred those that contained examples of both diary writing and filter-type writing. The distinction made between the two is that diary entries should refer to the author's experiences, while filter entries refer to events that do not directly involve the author. Although all of the journals collected for this study were of the diary type, we did not control for differences between individual diary-type and filter-type entries, and it is possible that male authors were more likely to include filter-type entries than women in this sample. There was little variation in pronoun use by age, although the oldest users (30+) tended towards using first person pronouns less frequently than the youngest users (<20). This might be taken to indicate a more self-referential style for teenage users than for older users. However, this result may be influenced by the smaller sample size available for older authors.

The category of "first person" clearly stratifies the corpora along a continuum Diary > Dem > Fic > News > Acad, correlating with the relative frequency of first person pronouns in these corpora. Relative frequency of all first person pronouns decreases steadily as one moves through this continuum: 77.82 tokens per 1000 words (Diary) > 54.72 (Dem) > 24.84 (Fic) > 9.54 (News) > 5.88 (Acad). The continuum corresponds, in a general way, to a degree of personal involvement in the text and it is not surprising that the Diary genre occupies one end of this continuum. The result confirms what is already known about the first person forms as they occur in the better known genres such as those represented in BNCBaby. Biber et al (2000, p. 333) note, for example: "With the exception of *we/us*, forms which refer to the speaker and the addressee (*I/me*, *you*) are far more common in conversation (and to a lesser extent fiction) than in other registers." Furthermore, it is the *I* form which dominates as the inflectional form in all genres. This fact is in line with Aarts' (2004, pp. 36-38) finding that that personal pronouns occur significantly more frequently in subject positions than in non-subjects positions in all categories of text.

To discuss this variation in greater detail, we will proceed by investigating each of the first person forms in their own right, rather than treating them as one class. Furthermore, we will examine the degree of variation within each corpus to ascertain the level of internal consistency within each genre.

Frequency of Singular Forms

The discussion above points to Dem and Fic as being the two genres which are closest to the diary corpora and hence most interesting to compare with these diary corpora. We proceeded, therefore, to explore the behavior of *I*, *me*, and *my* in more detail in the Diary, Dem, and Fic corpora. Figures 1-3 are "notched" boxplots of the relative frequencies of these forms in the UKDiary, USDiary, Dem and Fic corpora. The vertical length of a box corresponds to the amount of variability – the larger the box, the greater the spread of the data. The dark horizontal line in a box represents the median and the length of the box represents the difference between the 25th and 75th percentiles. Maximum and minimum values (apart from any outliers) are indicated by the extremes of the whiskers. Where the notches about two medians do not overlap, the medians are, roughly, significantly different at a 95% confidence level (cf. McGill, Tukey, & Larsen, 1978; Potter, 2006). The boxplots of *I* in Figure 1 suggests similar behaviour of *I* in UKDiary, USDiary and Dem, though the notches in the three boxes corresponding to these three sets of data do not obviously overlap, meaning that there would appear to be a significant difference between the three corpora with respect to this parameter. There is, of course, no overlap between the notches of any of these three datasets and that of Fic, confirming a significant difference between Fic and other corpora. Boxplots of *me* in Figure 2 show overlapping notches for UKDiary and USDiary, on the one hand, and overlapping notches for Dem and Fic on the other hand. Figure 2 shows clearly how the behavior of the *me* form separates out the diary corpora from the others. Figure 3 summarizes the distribution of *my* and suggests a division between the Diary corpora and the other genres represented by BNCBaby, similar to what is seen in Figure 2. However, the non-overlapping (or borderline overlapping) notches indicate significant differences for all comparisons of the corpora. Summing up, then, the boxplots show that the frequency data for *me*, more so than for *I* and *my*, is consistently similar in the diary corpora and most effectively differentiates the two diary corpora from both Dem and Fic.

Contextual Patterns of Singular Forms

Although the frequency of *I* in the Diary corpora is subject to some variation, its mean frequency in both Diary corpora is higher than in other genres. This is not surprising, given the overall expectation that in an online journal the author is primarily engaged with reflections on their own life. More specifically, the online diary style offers opportunities for authors to indulge in sustained and uninterrupted confessional outpourings which one would be unlikely to encounter in conversation. The excerpt in (1)–a continuous stretch of writing in one diary entry, including misspellings such as *sentances* for *sentence* –illustrates this kind of style.

(1) From UKDiary

Figure 1. Boxplots of I in four corpora

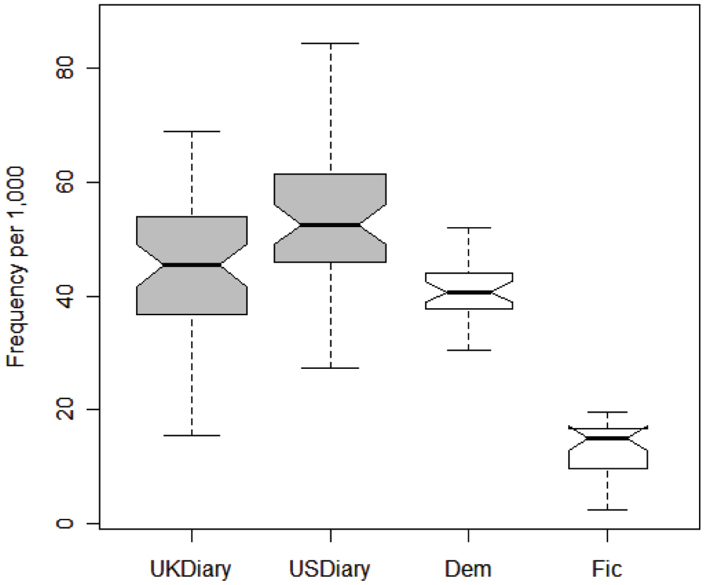
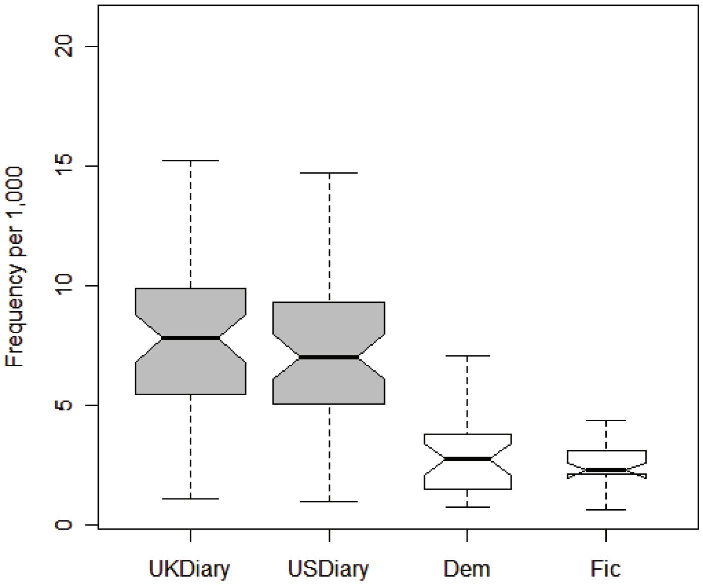


Figure 2. Boxplots of me in four corpora

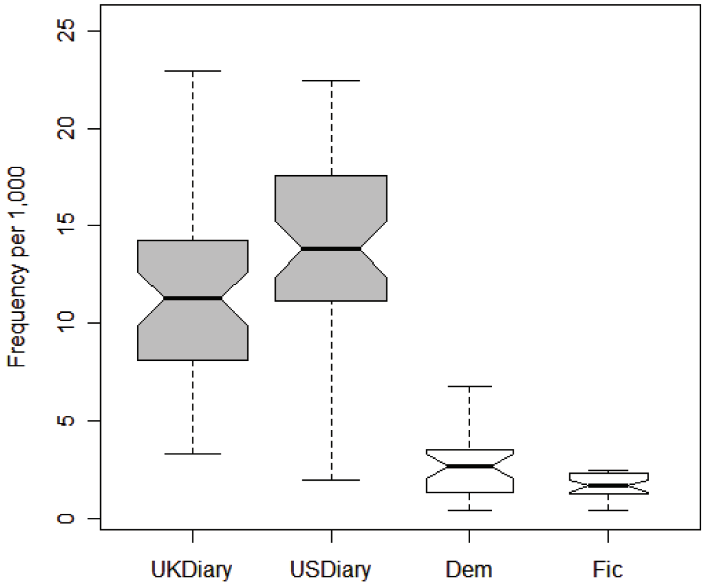


Who am I exactly?

I know.

I know who I want to be. And if you want to be something, then surely deep down, you ARE that person.

Figure 3. Boxplots of my in four corpora



I don't know how to talk properly, I can barely get my sentances out.

I stand there stuttering & repeating myself and by the time I say what I wanted to, the person is sure to have lost interest.

I spent a good part of my life shy, and I got over that. Now i'm back there again & i'm scared of the outside world.

I'm scared of being myself incase I get shouted at.

I'm scared to show people my writing because as soon as I proofread for their eyes my words are tangled & clumsy.

Since it is the *me* form which best differentiates diary corpora from the others, we decided to explore this form further by studying some of the contextual patterning with *me*. We selected UKDiary and Dem corpora for this purpose, since they are both based on British English. We used Wordsmith Tools (Scott, 2004) to retrieve and display collocates of *me* in the two corpora in a window of three words to the left and three to the right (L3-R3). In the case of the Dem corpus, which has a considerable amount of meta data in the XML markup, we made use of Wordsmith's options to ignore extraneous markup in calculating collocates and statistics.

Tables 2 and 3 show the top ten collocates for *me* in UKDiary and Dem respectively. The "Word" column contains the collocates, which are calculated around the search word, *me*. The collocates are sorted by descending strength of association (the "Relation" column), as computed by Wordsmith Tools using the log likelihood measure. One difference that can be seen immediately is that in Dem the top collocates are consistently concentrated to the left of the search word in the L1 position, whereas this is not the case for UKDiary. So, for example, in the UKDiary *to* occurs with equal frequency in the L2 and R1 positions as a collocate for *me*;

Table 2. Top 10 collocates of me in UKDiary sorted by log likelihood score for relation, calculated over L3-R3

Word	**Relation**	**Total**	**L3**	**L2**	**L1**	**R1**	**R2**	**R3**
to	574	197	23	**54**	36	**54**	15	15
and	368	153	26	26	6	**63**	16	17
for	260	80	5	7	**40**	13	6	9
you	210	66	20	**27**	0	4	11	4
let	174	24	1	0	**22**	0	1	0
told	154	21	2	0	**18**	0	1	0
he	145	39	8	**18**	0	3	6	4
that	135	56	12	**16**	0	13	5	10
with	129	48	1	1	**32**	6	6	2
a	112	77	9	2	0	**35**	15	16

and, as a collocate of *me*, is concentrated in the R1 position etc. The two displays thus point to a relatively diffuse kind of collocational behavior for *me* in the UKDiary.

Strong collocational relationships with *to*, *for*, *with*, *told* and *let*, all in the L1 position, can be observed in both Tables 2 and 3. *Told* and *let* in this position suggest that *me* is most likely functioning as the object of the verb and, simultaneously, as the subject of a following infinitival clause with the infinitive following the *to*, as in *Mum told me to visit her*. The prevalence of this construction serves as a reminder that the "object" form, *me*, cannot be equated with purely a patient (as opposed to agent) role in English. Indeed, the results in Tables 2 and 3 suggest that the agent role is a conspicuous and important semantic role for *me* in both corpora. *For* and *with*, also among the top ten collocates of both corpora, may also be functioning in the L1 position to introduce infinitival clauses, suggesting that *me* in at least some of these cases may be functioning as the understood subject of an infinitival clause.

Table 3. Top 10 collocates of me in BNCBaby Dem sorted by log likelihood score for Relation, calculated over L3-R3

Word	**Relation**	**Total**	**L3**	**L2**	**L1**	**R1**	**R2**	**R3**
to	1,167	485	0	0	**302**	175	8	0
told	730	109	0	0	**109**	0	0	0
for	609	213	0	0	**175**	33	5	0
give	579	115	0	0	**111**	4	0	0
let	543	89	0	0	**88**	1	0	0
tell	439	88	0	0	**87**	1	0	0
with	354	132	0	0	**115**	17	0	0
want	282	102	0	0	**88**	13	1	0
gave	177	29	0	0	**29**	0	0	0
help	170	33	0	0	**30**	3	0	0

A point of particular interest concerns the collocational patterning of *to* and *me* in these corpora. In the combination *to me*, *me* is clearly the object of the preposition *to*. As part of the combination *me to*, we would be expect that *me* is functioning as the understood subject of an infinitival clause, as discussed above, and as found in *Mum told me to visit her*. Table 4 shows 20 sample concordance lines from UKDiary illustrating the *me to* combination with the infinitival construction, including the unusual construction *she is proud of me to say…* in line 4. Note that the *me to* combination is more frequent than the *to me* combination in UKDiary (54 *me to* vs. 36 *to me*), but less frequent than *me to* in Dem (176 *me to* vs. 309 *to me*). It can be seen, then, that the infinitival clause construction with *me* as the understood subject is a feature of both UKDiary and Dem, though more associated with the former than the latter.

Frequency of Plural Forms

Plural forms of the first person pronouns are much fewer in number in the corpus and we will have less to say about these forms as a result. Once again, it is the subject form which dominates. Relative frequencies for first person plural *we* in the corpora are, in descending order: 9.416 tokens per 1,000 words (Dem) > 4.785 (Diary) > 3.45 (Acad) > 3.219 (Fic) > 2.643 (News). Here it is the Dem genre which evidences the highest use of *we*. This is quite different to what was found with the *I* form, where the Diary genre showed the highest relative frequency.[5] The results for *us* and *our*, on the other hand, show that these forms are indeed most frequent in the Diary corpora, similar to what was found with the corresponding singular forms.

Contextual Patterns of Plural Forms

It is not difficult to imagine why *we* is more frequent in Dem than in the diary corpora. Conversations provide continuous opportunities for jointly referencing the speech act participants, and for planning joint activities between them. Diary writing is different in both these respects. Even if online diary writing, as explained above, demands eventual "witnesses" for the writing, such witnesses are anonymous and virtual and not viable participants in future joint plans on the part of the author. One manifestation of the difference which can be easily quantified is the use of

Table 4. Sample concordance lines for me + to in UKDiary

And I have to text Wendy cos Mum told	me to	visit her. Lah. I'm gonna go read my book.
wasn't goimg to go, but instinct told	me to,	& I was glad I did. We had loads of fun
Anyway time for	me to	stop now because i need to go get milk
Shes really proud of	me, to	say that i was going through a pretty bad time
Mum really wants	me to	take Maths and maybe another science again
plot points that _peter had asked	me to	try and get across.
I want the doctor who sees	me to	be someone I am paying to listen to me
kept logging out unexpectedly, causing	me to	lose conversations and file transfers
An audience, waiting for	me to	slip & fall off this very tightrope.
to drag me out of my house and force	me to	be sociable, something I sorely need
LiveJournal is just too annoying for	me to	bother going through the whole thing
really stoned, and the one asked	me to	roll them a joint, i asked if they wanted it
What would you want	me to	say to you? It can be anything, but be honest

tag questions with *we*, as in *we don't have to go all the time, do we?* Tags with *we* have a natural place in conversation where they reference first and second person speech act participants. One would not expect them to be a feature to the same extent in the Diary corpora. Indeed, a search on *we?* (including *are we?*, *did we?*, *will we?*, *should we?* etc.) in Dem yielded 437 tokens, representing 4.58% of the total number of *we* forms in Dem. There is not a single instance of *we* used as a tag in this way in either UKDiary or USDiary, even without a question mark after the *we*. Although the *we* tag, by itself, does not account for the sizeable difference in relative frequencies between the Diary corpora and Dem, it is indicative of the different modes of writing in the two genres. In the Diary corpora, *we* refers either to the author and others in the author's own life, as the author reflects on past events, as in the examples in (2), or are impersonal uses of *we* applying to humankind, as in (3). The Diary corpora have relatively little of the 'inclusive' use of *we* which we find in Dem, illustrated in (4).

2. *we* = author and author's circle, from USDiary
   a. *For the first time ever, New Years was actually a "family holiday" as in we actually did something instead of the kids going to a friends house*
   b. *Then ryan johnson came over with my favorite juice and we all drank gin and juice while playing monopoly. it was such a good time.*
   c. *God and I were spending some time together and He said to me, "Son, while we are here, do you think I could catch up on some things? I've been pretty busy*
   d. *And so we sat down on the couch in my living room, he clicked on the TV and we began to watch some prayers.*
3. *we* = human beings, from USDiary
   a. *Sometimes we think we need others to help us and we become dependent on them*
   b. *Each action we take sustains a pattern or breaks it. Each word we speak reinforces things as they are, or moves toward change.*
   c. *It's that kidlike innocence that pulls us back into the real reality of it all: we can fight all we want, but we are only delaying the celebration.*
   d. *Others stay awhile and leave footprints on our hearts and we are never, ever, the same*. (cited as a quote within an entry)
4. *we* = author and addressee, from Dem
   a. *Shirley's sort of getting on to you a bit I think we'd better make a move.*
   b. *We're not there yet are we?*
   c. *Oh which way are we approaching it?*
   d. *We don't want to rush them eating it do we?*

Responses to Online Diary Entries

In Section 1, we referred to an unspoken expectation on the part of online diary writers that there will be a "witness to the confession" and that online responses and comments on such diaries are, in a sense, integral to the full enactment of online diary writing. It is of some interest, therefore, to study the use of pronouns in these responses. In particular, one may inquire as to whether pronoun use in responses to online diary entries mirrors the use of pronouns in the diaries themselves, or whether their use is more comparable to that found in Dem. More so than the diary itself, the responses might be expected to engage with an addressee (the diary author in the case of the response writer) and hence we might expect to find more prevalent use of the second person forms in responses than in the main diary texts.

In order to determine the nature of pronoun use in our diary responses, we revisited the online diaries making up the Diary corpora approximately one year after initial data collection. We collected all available comments posted in response to the diary entries recorded in the corpus. One year later, twenty-one of the original diaries were no longer available, eleven of those in UKDiary and ten in USDiary. The summed corpus of responses is imperfectly balanced, with 132,831 words recorded in the UKDiary responses and only 26,257 words recorded in the USDiary responses (159,088 in total). Overall, the average size of the comment log for each author was 2,376 words (UKDiary: 3,406, USDiary: 691). However, the UKDiary responses benefit from a prolific series of comments in a single blog resulting in over 60,000 words, and without that series of comments, the UKDiary average falls to 1,900 words and the overall average falls to 1,620. A final caveat to these data is that, given the nature of online communication, there is no guarantee that responders hailed from the same geographic location as authors. All comments are in English, but while subcorpora have been coded as "USComments" and "UKComments", referring to the location of the blog author, the location of the commenter is ultimately uncertain. It is unlikely, however, that all respondents were unknown to the diary author, given that many responses included references to emotional connections to the author (e.g., *I miss you*, *I love you*) and to sharing a physical location with the author (e.g., *it was good to see you again*).

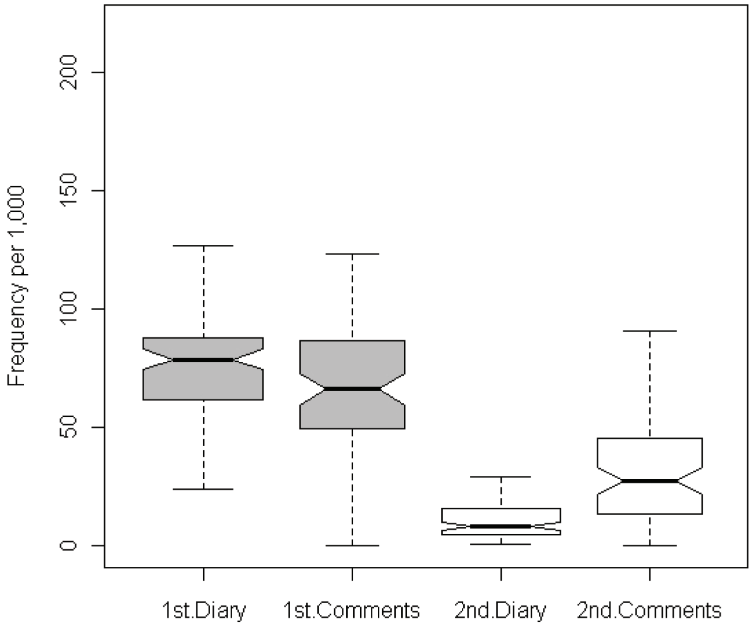
As in the Diary corpora, first person singular forms are the most commonly occurring pronouns recorded within the comments. *I* occurred in the comments about as often as it occurred in UKDiary and USDiary, if slightly less frequently. However, there was great variation recorded in USComments, where first person singular pronouns were sometimes very frequent in the responses to a single diary, and sometimes completely unattested. This high variability may be a reflection of the smaller sample size available for USComments.

Returning for a moment to UKDiary and Dem, recall that *me* acted as the understood subject of infinitival clauses in *me to*, and that this behaviour was relatively more frequent in UKDiary than in Dem when compared to the use of *me* as an object following a preposition in *to me*. If we examine UKComments, we find that it patterns more closely with Dem (*me to*: 14 vs. *to me:* 26). Here, it appears as if the writer and responder are engaged in a dialogue that more closely approximates conversation than is observed in diary text.

My occurred more frequently in the Diary corpora than in the comments for both UK and US regions, and one might imagine that this difference is due to the relative degree of self-reference in the text. While a diary entry is, by its nature, self-referential, the comments collected were written as responses and not necessarily as equivalent personal commentaries on the same topic. First person plural forms patterned similarly to their UKDiary and USDiary counterparts, although there were no recorded instances of *our* in USComments.

Following *I*, *you* was the most frequent pronoun recorded in UKComments and USComments. We find that in the recorded comments, *you* is used significantly more frequently than it is used in the main diary text (Figure 4). This is perhaps unsurprising, as comments are directed towards the original author and the content of his or her diary entry. Common word clusters include *I miss you* (11), *I think you* (11), *if you want to* (9), *if you don't* (9), *you have to* (9), *agree with you* (6), *hope you feel better* (6), and *you could always* (5). Responses generally referred back to the original diary entries, or to a thread present in the comments. For example, a response such as *hope you feel better* relates directly to posted material when an author tells his audience that he has been unwell, while phrases like *I (so, absolutely, heartily) agree with you* express solidarity with the author and his or her opinions, stated immediately previously. In some cases, particularly

Figure 4. Boxplots of first and second person in Diary and Comments corpora



for *I miss you*, this sentiment was expressed both by the commenter and by the diary author, in a secondary response to the commenter.

In both the original diary corpora and in the comments corpora, the first person is more commonly observed than the second person. However, in our sample of responses, the second person is used more often in both UKComments and USComments than in either UKDiary or USDiary. The frequency of *you* in comments reflects the role of feedback to authors, and to a certain extent, a type of dialogue between the original author and his audience that is not so completely dissimilar from conversation.

FUTURE TRENDS

Throughout this study, we have compared data from recognized genres, such as fiction and conversational speech, to a genre that is growing up out of communicative opportunities available on the Internet. Through displayed patterns of pronoun use, we can see that characteristics displayed in online diaries mark this as a genre in its own right. The online diary model allows, and as we have discussed, may indeed *require* an audience. While entries might not be specifically directed towards an audience, an audience is, at the very least, implicit in the diary's availability online. Along the dimension of pronoun use, responses to online diaries are not dissimilar to spoken conversation. Online diaries, then, can be considered to be one of several emergent forms of communication that are made possible by a digital infrastructure, and given their popularity, are a thriving component of the new media. Expanded research in this area might focus on determining further characteristics of this and other internet-based genres, such as email, or other blog types (e.g., political versus personal). Such research may also help to compare the types of communication between correspondents. Finally, from our perspective, our results further the notion that the study of inflectional forms of words and categories is vital for greater understanding of language use. Future studies that focus on inflected forms, both of categories (e.g., *I* for pronouns) and of words (e.g., *remembers* as a part of *remember*), will be able to identify patterns of use that go beyond the level of word lemma. In this way, we can generate a more complete description of language use across genres, and in turn, increase our understanding of language use in a digital communicative world.

CONCLUSION

Clearly, the individual inflectional forms of first person pronouns have different roles to play in distinguishing the genres discussed here. In keeping with the "personal involvement" character of online diary writing, the singular first person forms are more frequent in the Diary corpora than in other genres. Even so, there is considerable fluctuation in the use of these pronoun forms within the Diary corpora. While the total number of *I* forms in the diary corpora exceeds that of the other corpora, it is the *me* form which shows the most consistent behaviour within the UKDiary and USDiary and which most effectively differentiates the diary corpora from other genres. Partly, these results reflect a confessional style of expression which is not found to the same degree in the other genres. The use of *me* as the understood agent of a following infinitival clause seems to be relatively more common in the Diary corpora, compared with, say, the conversational corpus. Of the first person plural forms, it is *we* which behaves most consistently within the Diary corpora and differentiating these corpora from the other genres. Diary writing does not employ the 'inclusive' use of *we* to the same extent as is done in conversation and this is one factor leading to the higher frequency of *we* in conversation compared with diary writing.

REFERENCES

Aarts, F. G. A. M. (2004). On the distribution of noun-phrase types in English clause-structure. In Sampson, G., & McCarthy, D. (Eds.), *Corpus linguistics: Readings in a widening discipline* (pp. 35–48).

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

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (2000). *Longman grammar of spoken and written English*. Harlow, UK: Pearson Education Limited.

Constant & A. Dister (Ed.). Actes du 26e Colloque international Lexique Grammaire (pp. 161-166).

Haegeman, L., & Ihsane, T. (1999). Subject ellipsis in embedded clauses in English. *English Language and Linguistics*, *3*(1), 117–145. doi:10.1017/S1360674399000155

Haegeman, L., & Ihsane, T. (2001). Adult null subjects in the non-pro-drop languages: Two diary dialects. *Language Acquisition*, *9*(4), 329–346. doi:10.1207/S15327817LA0904_03

Herring, S. C., & Paolillo, J. C. (2006). Gender and genre variation in weblogs. *Journal of Sociolinguistics*, *10*(4), 439–459. doi:10.1111/j.1467-9841.2006.00287.x

Lee, D. (2001). Genres, registers, text types, domains, and styles: Clarifying the concepts and navigating a path through the BNC jungle. *Language Learning & Technology*, *5*(3), 37–72. Retrieved from http://llt.msu.edu/vol5num3/lee/.

Lenhart, A., & Fox, S. (2006). *Bloggers. A portrait of the internet's new storytellers*. Washington, DC: Pew Internet & American Life Project. Retrieved February 12, 2007, from http://www.pewinternet.org/pdfs/PIP%20Bloggers%20Report%20July%2019%202006.pdf

London & New York. Continuum. (Reprinted from Lingua, 26, pp. 281-293, 1971).

Marne-la-Vallée, France: Université de Marne-la-Vallée, Institut Gaspard-Monge.

McGill, R., Tukey, J. W., & Larsen, W. A. (1978). Variations of Box Plots. *The American Statistician*, *32*, 12–16. doi:10.2307/2683468

McNeill, L. (2003). Teaching an old genre new tricks: The diary on the internet. *Biography*, *26*(1), 24–47. doi:10.1353/bio.2003.0028

McNeill, L. (2005). Genre under construction: The diary on the internet. *Language@Internet, 2*, article 1. Retrieved January 28, 2009, from http://www.languageatinternet.de/articles/2005/120

Miura, A., & Yamashita, K. (2007). Psychological and social influences on blog writing: An online survey of blog authors in Japan. *Journal of Computer-Mediated Communication*, *12*, 1452–1471. doi:10.1111/j.1083-6101.2007.00381.x

Newman, J. (in press). Balancing acts: Empirical pursuits in cognitive linguistics. In Glynn, D., & Fischer, K. (Eds.), *Quantitative methods in cognitive semantics*. Berlin: Mouton de Gruyter.

Newman, J., & Rice, S. (2004). Patterns of usage for English SIT, STAND, and LIE: A cognitively inspired exploration in corpus linguistics. *Cognitive Linguistics*, *15*, 351–396. doi:10.1515/cogl.2004.013

Potter, K. (2006). Methods for presenting statistical information: The Box Plot. In H. Hagen, A. Kerren & P. Dannenmann (eds.), *Visualization of large and unstructured data sets*, (LNI Vol. S-4, pp. 97-106). Bonn, Germany: Gesellschaft für Informatik (GI). Retrieved from http://www.cs.utah.edu/~kpotter/pubs/IRTG2006kpotter.pdf

Salber, C. (2001). Bridget Jones and Mark Darcy: Art Imitating Art…Imitating Art. *Persuasions on-line, 22*(1). Retrieved December 16, 2008, from http://bridgetarchive.altervista.org/art_imitating_art.htm

Scheibman, J. (2001). Local patterns of subjectivity in person and verb type in American English conversation. In Bybee, J., & Hopper, P. (Eds.), *Frequency and the emergence of linguistic structure* (pp. 61–89). Amsterdam, Philadelphia: John Benjamins.

Scott, M. (2004). *WordSmith Tools, Version 4*. Oxford, UK: Oxford University Press.

Stefanone, M. A., & Jang, C.-Y. (2008). Writing for friends and family: The interpersonal nature of blogs. *Journal of Computer-Mediated Communication*, *13*, 123–140. doi:10.1111/j.1083-6101.2007.00389.x

Stubbs, M. (2001). *Words and phrases: Corpus studies of lexical semantics*. Oxford, UK: Blackwell.

Tao, H. (2001). Discovering the usual with corpora: The case of *remember*. In Simpson, R., & Swales, J. (Eds.), *Corpus linguistics in North America: Selections from the 1999 symposium* (pp. 116–144). Ann Arbor, MI: University of Michigan Press.

Tao, H. (2003). A usage-based approach to argument structure. *International Journal of Corpus Linguistics*, *8*, 75–95. doi:10.1075/ijcl.8.1.04tao

Teddiman, L., & Newman, J. (2007). Construction of and findings from a diary corpus. In Camugli, C. (Ed.), *M*.

KEY TERMS AND DEFINITIONS

Boxplot, Box Plot: A standard graphical visualization of numerical variation in data, including 5 key statistics: median, 25th and 75th percentiles, and maximum and minimum values present in the data. “Notched” boxplots allow quick inspection of significant differences in variation between two or more datasets. Where the notches do not overlap, there are significant differences between the datasets.

Collocate: Collocates of word *x* are those words that occur in the environment of *x*, within a text. For example, in the sentence *It is time for me to go*, *go* is a collocate of *me* at position R2 (second word to the right of *me*).

Corpus: A collection of written texts or transcriptions of spoken language. Now understood to be an electronic collection.

First Person: In English linguistics, first person refers to the pronouns *I, me, my, mine, myself* (singular), *we, us, our, ours, ourselves* (plural).

Genre: A group of texts collected for corpus-based studies. Typically, collected texts are drawn from a cohesive domain, e.g., press, religion, fiction, academic, private letters, and diaries.

Lemma: A representation of a word that subsumes all its inflected forms. For example, the lemma verb sing includes the inflected verb forms *sing, sings, singing, sang, sung*.

Online Diary: A type of weblog (blog) that is used by the author as a personal journal but is publicly available for others to view and comment upon. Differs from other blog types in that the subject matter is grounded in the experiences of the author, and is not thematically based.

Second Person: In English linguistics, second person refers to the pronouns *you, your, yourself, yourselves, yours*.

ENDNOTES

[1] Stubbs (2001, p. 99) draws attention, in passing, to the issue of investigating different inflected forms, as opposed to lemmas though the idea is not further explored.

[2] One website offering advice on writing diaries recommends, along with many other tips: "Think of a diary as a conversation with someone. When read, the words sound like someone talking to you. Think of sharing your thoughts about when and where the event took place, how you felt about it, "gossip" or comment about other people and so on. Look at the situation from different angles." (Power of the Real World website http://english.unitecnology.ac.nz/resources/units/real_world/diary.html) As far as pronoun usage is concerned, the same website advises that diaries are "written in first person I".

[3] "I shamelessly stole the plot from *Pride and Prejudice* for the first book. I thought it had been very well market-researched over a number of centuries and she probably wouldn't mind" (words attributed to Helen Fielding, Daily Telegraph 11/20/1999, cited by Salber, 2001). Hence, the title of Salber's article: "Bridget Jones and Mark Darcy: Art Imitating Art…Imitating Art".

[4] McNeill (2003, p. 28) reflects on a possible genre distinction between "online journal" and "weblog", with online journals being more meditative and processed and weblogs being more immediate and "off-the-cuff". McNeill does not accept any strict separation of the two, noting: "In reality, though, even the scantiest of blog narratives incorporates 'trademark' diary features, with regular, dated, entries that focus on the diarist/narrator's experiences and interests." (McNeill, 2003, p. 29).

[5] This result does not accord with Biber et al's (2000, p. 333) observation, cited above, that the *we* form is not far more common in conversation than other genres. In the corpora studied here, *we* is approximately twice as frequent in Dem than it is in the other genres.