Language and Personality in Computer-Mediated Communication: A cross-genre comparison

Alastair J. Gill [a,*] Scott Nowson [b] & Jon Oberlander [b]

[a] *LEAD-CNRS UMR 5022, Université de Bourgogne,*
Pôle AAFE, Esplanade Erasme, BP 26513, 21065 Dijon Cedex, France

[b] *School of Informatics, University of Edinburgh,*
2 Buccleuch Place, Edinburgh, EH8 9LW, UK

Running Head: Language and personality in CMC

Total Wordcount: 7,215

[*] Please address queries about this work to Alastair J. Gill
Email addresses: `A.Gill@ed.ac.uk` (Alastair J. Gill), `S.Nowson@ed.ac.uk` (Scott Nowson), `J.Oberlander@ed.ac.uk` (Jon Oberlander).

Preprint submitted to JCMC *30 June 2006*

Abstract

It is known that personality is important in computer-mediated communication (CMC), influencing both how we express ourselves, and how we are perceived. Here we build in two ways on previous work which has used the LIWC text analysis tool to derive language factors relating to personality. First, we investigate whether linguistic factor structure in e-mail is similar to that in weblogs, and how these structures compare to previous findings for non-CMC, written language. Our findings broadly replicate results from offline studies, although blogs can be differentiated from non-CMC language, with e-mail sharing features with both genres. Secondly, we compare how the CMC language factors relate to personality, and seek differences between EPQ-R and Five-Factor Measures. We find that patterns of language behaviour for Neuroticism and Agreeableness distinguish CMC from non-CMC environments, while results from the two personality measures are largely compatible.

Introduction

Electronic media are pervasive, offering individuals many emerging channels for everyday communication (Baron, 1998; Crowston and Williams, 2000). For instance, we can reach the outside world by writing e-mails, running websites, or keeping weblogs. However, it appears that in each of these cases computer-mediated communication (CMC) can allow people access to information about fundamental aspects of our selves: our personalities. The lack of rules and expectations concerning CMC genres leaves a great deal of room for individual expression. Recent research has shown that personality has significant influence on the language patterns in e-mails or weblogs (Oberlander and Gill, 2006; Nowson, Oberlander, and Gill, 2005), and that strangers are able to make accurate personality judgements based on viewing an individual's website or e-mail (Vazire and Gosling, 2004; Gill, Oberlander, and Austin, 2006).

Given that personality can be projected and perceived in computer-mediated environments, we here explore a small set of questions. First, CMC may affect how people express themselves in general. So we consider whether the factor structure of linguistic patterns in e-mail is similar to that in weblogs, and how these factor structures compare to those found in written non-CMC. Secondly, different people may be affected by CMC in different ways. So we consider whether the language factors relate to personality differences in the same way as in non-CMC. Pursuing this second question, we compare how the language factors relate to three- and five-factor models of personality, with the expectation that the model with fewer personality factors could help find stronger correlations between personality and language use. To address these questions, we adopt the LIWC text analysis method developed by Pennebaker and colleagues (Pennebaker and Francis, 1999), since this allows direct comparison with self-disclosure in the form of diaries and other texts as reported in Pennebaker and King (1999).

The paper is structured as follows. In the next section, we review the diversity of CMC and compare it to traditional forms of speech and writing. We introduce the two main trait-based personality models, and the content analysis method which we will adopt in this paper. We also review previous literature which has examined linguistic characteristics of personality in writing, speech and CMC environments. On the basis of this previous work, we then re-state more precisely the research questions to be addressed. The following section indicates how we gathered two corpora to help examine language in e-mail and weblog computer-mediated environments. We then report two experiments in turn. The first compares the linguistic factor structure of the two types of CMC, and non-CMC writing. The second compares the relationships between personality and these language factors in CMC and non-CMC environments. In the latter case, the particular focus is on Extraversion and Neuroticism, but a secondary question concerns the role of Openness, Agreeableness and Conscientiousness versus Psychoticism. We finish up by discussing our findings and some of their implications.

Background

CMC and language

Computer-mediated communication, like other forms of writing, is less rich than face-to-face communication (Panteli, 2002). As a result, alternative linguistic strategies in e-mail and internet relay chat have been adopted to provide paralinguistic or social cues (Werry, 1996; Hancock and Dunham, 2001b; Colley and Todd, 2002).

Internet-based CMC should not be treated as a single genre, and is in fact composed of a number of distinct types of communication (Yates, 1996). For example, static webpages are for the most part wholly written, but instant messengers create (written) conversations that mimic spoken forms in many ways. Relatively stable varieties of internet CMC are emerging, although there is still variation within them, and new forms are continually evolving (Cho, 1996; Gruber, 2000). However, this has not prevented researchers from attempting to classify these forms of communication, either functionally or linguistically (Crowston and Williams, 2000; Shepherd, Watters, and Kennedy, 2004; Santini, 2005; Biber, 2004, 1988).

E-mail is one of the major contributors to the popularity of the internet. It has been estimated that 90% of internet users access e-mail (Fallows, 2005), while 73% of American adults access the internet (Madden, 2006), suggesting that over 130 million people use e-mail in the US alone. It is a written form, in which interlocutors are physically separated; it is also durable, and authors often use complex linguistic constructions; however, e-mail is often unedited, uses first- and second-person pronouns, present tense and contractions, and it is generally informal in tone (Bälter, 1998; Baron, 2001). Indeed, characteristic e-mail features have been identified, including ellipses, capitalisation, extensive use of exclamation marks, and question marks (termed 'e-mailism'; Colley and Todd, 2002). As a result, e-mail is often considered intermediate in form between speech and writing (Yates, 1996; Baron, 1998; Gruber, 2000; Nowson *et al.*, 2005, cf. Collot and Belmore, 1996). However, e-mail certainly differs from speech, being more verbose, yet less emotional (Whittaker, 2003).

If e-mail use statistics are hard to calculate, weblog numbers are almost impossible. A weblog is a frequently updated website which contains news and views on a variety of topics, from politics to gossip, and weblogs have been regarded as a powerful news-gathering tool (Belo, 2004). Quantitative studies rarely take into consideration non-English language weblogs, nor do they take account of inactive weblogs or authors with multiple sites. However, by way of an illustration of their increasing popularity, it has been estimated that by the end of 2005, 53.4 million blogs would have been created, using just the leading blog providers (Henning, 2005). The term ‘blog’ is perhaps more widely used than ‘weblog’, and it is most commonly used to refer to the sub-category of online personal diaries; blogs are the focus of what follows. In contrast to traditional diary-keeping, blogging can also be a social activity (Marlow, 2004; Efimova and de Moor, 2005).

Like e-mail, blogs can also demonstrate characteristic linguistic features, with posts written in ‘short, paratactic sentences’ employing ‘informal, non-standard constructions and slang’ (Nilsson, 2003). However, due to the social nature of blogging, it is also possible to observe communities and social groups, including in-group and out-group language behaviour (such as, *I*, *me*, *my*, *we*, *us* and *our*, rather than *they*, *them* and *their*) and shared background knowledge and concepts (Nilsson, 2003; Cassell and Tversky, 2005, cf. Brown and Yule, 1983), as well as shared responses to traumatic events such as September 11, 2001 (Huffaker, 2004; Cohn, Mehl, and Pennbaker, 2004; Krishnamurthy, 2002).

Models of Personality

In this paper we refer to two main models and associated measurements of personality: Eysenck's three-factor model (Eysenck and Eysenck, 1991; Eysenck, Eysenck, and Barrett, 1985), and the five-factor model (Digman, 1990; Costa and McCrae, 1992; Wiggins and Pincus, 1992; Goldberg, 1993). Both of these describe Extraversion (Extraversion–Introversion) and Neuroticism (Emotionality–Stability) which are undisputed and central to theories of personality. To these, the three-factor model adds the trait of Psychoticism, while the five-factor model adds Openness, Agreeableness and Conscientiousness (Matthews, Deary, and Whiteman, 2003; Lippa and Dietz, 2000).

Very roughly speaking, Extraversion measures how sociable or energetic someone is; Neuroticism measures how anxious or worrying there are; Psychoticism measures how tough-minded or individualistic they are; Openness measures how intellectual or open-minded they are; Agreeableness measures how good-natured or co-operative they are; and Conscientiousness measures how orderly or dependable they are.

Although there are theoretical differences between models (Deary and Matthews, 1993; Eysenck, 1970; Eysenck and Eysenck, 1991; Eysenck, 1993; Block, 1995; Matthews *et al.*, 2003; McCrae and Costa, 1987, 1997; Funder, 2001; Buss and Finn, 1987; Pytlik Zillig, Hemmenover, and Dienstbier, 2002), a prominent question is how these models relate to each other: Kline (1993) notes that for Eysenck's EPQ measure, 'Extraversion and Neuroticism are clearly identical to two of the big five factors and Psychoticism would appear to be a mixture [of the other three traits]'. In discussing our findings from both models, we consider this relationship later in the paper.

Content analysis and factor analysis

Content analysis focusses on context-independent occurrences of lexical content words in written text. Although there are many different approaches (see Mehl, 2005; Pennebaker, Mehl, and Niederhoffer, 2003; Smith, 1992, for an overview; and Oberlander and Gill, 2006 for alternative approaches), here we describe one method in particular, which uses the Linguistic Inquiry and Word Count text analysis program (LIWC; Pennebaker and Francis, 1999). [1]

Building on earlier work by Gottschalk and Gleser (1969), LIWC counts occurrences of words or word-stems belonging to pre-defined semantic and syntactic categories (which belong to four main groups: Linguistic Dimensions, Psychological Processes, Relativity, and Personal Concerns). These provide different ways of describing texts (Kilgarriff, 2001; Lowe, 2004, cf. semantic tagging Rayson, 2003). For instance, using this system, words like *could*, *should* and *would* are categorised as 'discrepancies', allowing the overall percentage of 'discrepancy' words to be calculated for the text as a whole. One of the major strengths of the LIWC text analysis approach is that the analysis program and the dictionaries have been rated and validated by independent judges (Pennebaker and Francis, 1999; Pennebaker and King, 1999). Although LIWC counts some syntactic features, such as pronouns, and verbs of various

[1] Note that more recent versions of the program have been released (LIWC2001; Pennebaker, Francis, and Booth, 2001, also the forthcoming LIWC2006), but to ensure comparability with results obtained using LIWC, we describe this version here.

tenses, these are not derived from a part-of-speech analysis of the data (cf. Gill, 2004; Oberlander and Gill, 2006).

Originally used to examine the relationship between language use in disclosure with measures of health and well-being (Pennebaker, Mayne, and Francis, 1997; Pennebaker, 1997; Graybeal, Sexton, and Pennebaker, 2002; cf. Oxman, Rosenberg, Schnurr, and Tucker, 1988), LIWC has since been used to study a number of linguistic behaviours, including deception (Newman, Pennebaker, Berry, and Richards, ress), gender (Mehl and Pennebaker, 2003), emotional tone (in newsgroups, Joyce and Kraut, 2006), add JCMC paper using LIWC, here, and individual differences (Pennebaker and King, 1999). We now discuss the last of these studies in greater detail, because it provides a reference point for comparisons with language use and personality in CMC.

Personality language

Pennebaker and King (1999) analysed texts written by authors for whom five-factor personality information was available. The studies used multiple writing samples produced by over 800 participants of undergraduate level summer schools. Factor analysis was used to derive a small number of factors grouping individual LIWC variables and these were then correlated with writers' scores on personality dimensions. We note the similarity of this approach to that which Biber used to explore genre (*e.g.*, Biber, 1995). The four derived factors were: Making Distinctions, Immediacy, Social Past, and Rationalization. Personality dimensions related to the factors as follows: Extraverts used language associated with the Social Past, and avoided language associated with Making Distinctions; Neurotic individuals used language associated with Immediacy;

Individuals scoring high in Openness used language associated with the Social Past (and some related to Making Distinctions), and avoided language associated with Immediacy; High Agreeableness scorers used language associated with Immediacy; High Conscientiousness scorers avoided language associated with Making Distinctions.

In a recent study of personality and language, Mehl, Gosling, and Pennebaker (2006) used LIWC to analyse speech sampled from everyday interaction. Overall, they found that Extraverts have a higher word count, with shorter words; Neurotics have a lower word count; people with higher levels of Openness talk less about social processes, use fewer past tense words and third-person pronouns; individuals higher in Agreeableness use more first-person pronouns, fewer articles and fewer swear words; and those higher in Conscientiousness use fewer words relating to negative emotions and swearing.

Analysis using LIWC is not the only work to examine the influence of personality upon language and interaction, but most of the rest has focused on Extraversion. For instance, it has been found that Extraverts: initiate more laughter; express more pleasure talk, agreement, and compliments; use more self-referent statements; and talk more, focusing on extra-curricular activities. Introverts, on the other hand, use more language relating to hedges and problem talk (Gifford and Hine, 1994; Thorne, 1987). Additionally, Extraverts have been shown to talk more in at least some situations (Carment, Miles, and Cervin, 1965, cf. Thorne, 1987).

At a lexico-grammatical level, Extraverts use a greater proportion of pronouns, adverbs, verbs and a higher total number of words in formal and informal situations, with Introvert language features more closely related to formal language (Furnham, 1990; Dewaele and Furnham, 1999, 2000; Dewaele, 2001). For reviews of personality and speech features, see also Scherer (1979) and Smith (1992).

Considering CMC specifically, it has been argued that its perceived anonymity allows people to feel more comfortable participating in interactions, where they may feel less able to do so in face-to-face (FTF) environments (Bloch, 2002; Yellen, Winniford, and Sanford, 1995), or even be more likely to engage in deceptive behaviour (Hancock, Curry, Goorha, and Woodworth, 2006). The liberating effect of CMC can also be noted in the intimacy of topics discussed in weblog and e-mail contexts, and perhaps even more notably in online personal advertisements (Groom and Pennebaker, 2005), although we still suppose that audience effects are present (Bell, 1984), and mediate the content relative to a traditional, private (non-online) diaries. Indeed, ratings of personal websites by unacquainted individuals showed evidence of enhancement in their projection of Extraversion and Agreeableness. However, this form of asynchronous CMC still demonstrates good target-judge agreement, especially in the case of Openness (Vazire and Gosling, 2004, although cf. Hancock and Dunham, 2001a for some issues regarding personality perception in CMC).

Similarly, studies of language in CMC have demonstrated that there are systematic patterns associated with e-mail and weblog data. In weblogs, authors' use of individual grammatical categories (parts-of-speech; POS) has been found to correlate with levels of their Openness and Agreeableness. More Agreeable authors use more articles, and fewer verbs, adverbs and interjections; authors with higher levels of Openness use more adjectives and prepositions and, like the more agreeable authors, fewer adverbs (Nowson *et al.*, 2005; Nowson, 2006). A more sophisticated analysis of an e-mail corpus has identified multi-item patterns of words and word-stems, and parts-of-speech, in relation to personality (Gill and Oberlander, 2002; Oberlander and Gill, 2006). Bottom-up stratified corpus comparison, showed, for instance, that High Extraverts used collocations involving inclusive expressions and connectives more broadly, generating conjoined noun phrases, and collocations involving proper names.

Research Questions

With this background in place, we can state more precisely the questions addrtessed in what follows:

(1) (a) Is the linguistic patterning of factor structure in e-mail similar to that in weblogs?

(b) How do their structures compare to those found in written non-CMC?

(2) (a) Do these factors relate to personality differences (primarily Extraversion and Neuroticism) in the same way as in non-CMC?

(b) Are any relations stronger or weaker if Psychoticism is substituted for the Agreeableness, Conscientiousness and Openness of previous studies?

Method

To tackle these questions, we require corpora containing examples of two subtypes of CMC: e-mail, and blogs. We describe in turn the data collection methods used to construct them.

Corpus 1: E-mail

Participants

105 current or recently graduated university students took part in this study . All participants were recruited via e-mail sent by the experimenter; they were not remunerated for their participation.

A sociobiographical questionnaire and Eysenck Personality Questionnaire-Revised (EPQ-R short version; Eysenck *et al.*, 1985) were administered to give information about the participants' backgrounds and personalities. 37 were males, and 68 females. The mean age of participants was 24.3. 53 were studying (or had studied) at an undergraduate level, and 52 at a postgraduate level.

All spoke English as their first language: 95 were of UK/Irish origin; 7 North American; and 3 Australasian. Scores on the personality dimensions were as follows (all scores between 0 and 12): Neuroticism (M=5.51, SD 3.2); Extraversion (M=7.91, SD 3.3); Psychoticism (M=2.90, SD 1.7); and Lie Scale (M=3.48, SD 2.2).

Materials and Procedure

The experiment was conducted on-line via an HTML form which participants filled in and then submitted over the internet. The web page had a simple design: After collecting the sociobiographical and personality information, as noted above, the participants went on to the tasks which were to write an e-mail to a good friend that 'you haven't seen for quite some time' describing

(a) in the first task what they had done in the past week, and (b) in the second task what their plans were for the coming week. Participants were advised to spend about 10 minutes on each task, write in their normal English prose, and were assured of confidentiality (but that they could feel free to substitute names of people or places if they desired).

Preparation of the corpus

This task generated 2 e-mail texts from each participant, giving 210 texts, and a total of around 65,000 words. Pre-editing was kept to a minimum to retain as much individuality as possible (for example, nonstandard words and spellings to imitate sounds), since these are characteristic of e-mail (Baron, 1998; Colley & Todd, 2002). Therefore, during a basic spell-check (using the standard Emacs spell-checker; Stallman, 1994) distinction was made between such intentional nonstandard spellings for communicative effect, and spelling errors.

Corpus 2: Blogs

Participants

71 authors of personal weblogs ('bloggers') contributed to this corpus. They were recruited by e-mail sent by the experimenter, or by word-of-mouth recommendation from fellow bloggers. To aid further participation bloggers were requested to include a link to the experimentation site in their weblog. No renumeration was given for taking part in this study.

As with the e-mail corpus, a sociobiographical questionnaire was administered online to participants, along with a personality questionnaire. This gave the following information about the participants: 24 were male, 47 female. Mean age=28.4, SD 8.3; 46 were educated to at least undergraduate degree level; 24 were of UK origin, 47 North American and 3 Australasian. For this corpus, an online implimentation of the IPIP Five Factor Personality Inventory (Buchanan, 2001) was used (rather than the EPQ-R), which provided the following information for the personality dimensions: Neuroticism (M=22.4, SD 6.3, out of a maximum score of 8), Extraversion (M=30.5, SD 6.5, maximum 9), Openness (M=29.3, SD 4.7, maximum 7), Agreeableness (M=26.3, SD 3.7, maximum 7) and Conscientiousness (M=31.8, SD 6.1, maximum 10).

Materials and Procedure

The experiment was conducted on-line via an HTML form for the presentation and submission of materials. Again the web page had a simple design. After an introduction to the experiment and assurances of confidentiality, the sociobiographical and personality questionnaires were presented. In the final stage, participants were requested to submit one month's worth of prior weblog postings. The month was pre-specified so as to remove the possibility of an individual choosing what they considered their 'best' or 'preferred' month, which may not be entirely representative.

Preparation of the corpus

Each participant provided one whole month of text from their blog. The blog corpus was then annotated using XML to mark high-level text features, such as content which could be described as 'Personal' and 'Commentary'. As in the e-mail corpus, stylistic editing of text was kept to a minimum in order to retain as much individuality as possible (for example, non-standard words and informalism). A basic spell-check (Wintertree, 2000) then corrected errors and standardised all spellings. Finally, text tagged as 'Personal' (i.e., written by the blog authors themselves) was extracted for analysis. This gave a corpus of 71 blogs, consisting of 1,854 individual posts (M=26.1, SD=20.3) and 411,843 words (M=5,801, SD=5,829). Word counts revealed that text of a personal nature accounted for 86.8% (SD 15.5%) of all the author-written (i.e., non-quoted) words.

Experiment 1: CMC factor structure

Previously, Biber (1988) used factor analysis of language use to distinguish writing styles across genres, but Pennebaker & King (1999) used it to study structure within comparable texts. Using a large corpus of student essays that had been passed through the LIWC tool (Pennebaker & Francis, 1999), Pennebaker & King chose 15 variables to use in their analysis (see section 1, below, for details of their selection criteria). In this experiment, we apply Pennebaker & King's factor analysis method to our e-mail and blog CMC data. In Experiment 2 (section 6) we go on to relate these factors to personality.

Analytic method

Scores on each of the LIWC's categories were required for each author. Following Pennebaker & King, each subject's scores are an average of the scores for each of their pieces of writing. For this, all 210 e-mail texts, and 1,854 individual personal blog texts were analysed with LIWC, and mean scores calculated for each subject.

In their study, Pennebaker and King (1999) outlined a number of considerations for selecting which of the 72 LIWC variables would be retained for factor analysis. Firstly, variables were retained from earlier validation studies only if they showed reliability of .60 or greater. Secondly, categories were required not to overlap. For example, the prepositions category was not included, since many inclusive and exclusive words are prepositions. Thirdly, categories were excluded if they did not refer to meanings or features of specific words (for example, word count). Similarly, current concern words (for example, *school*, *cash*, *ache*), with their topic-specific nature were also excluded. Finally, only variables that had a mean usage level of at least 1% were included.

The 15 variables to be included by Pennebaker & King in their factor analysis are listed in Table 1, along with mean frequencies for these variables within the e-mail and blog corpora, and those for written data reported by Pennebaker & King (1999). We now briefly describe the comparability of these data sets, before discussing the suitability of our data for entry into exploratory factor analysis.

There appears to be a basic underlying pattern: in fact, the rank of means from the blog and e-mail corpora are almost identical, bar the slightly higher frequency of Articles in blogs. Across all three studies, no variable is placed more than three places higher or lower; the difference is mostly only one place.

Perhaps the biggest differences are as follows. The blog corpus contains a greater frequency of words of length greater than six letters, but fewer examples of the present tense. The original written corpus contains more first-person

Table 1

Means (and ranks) of 15 LIWC variable scores for three studies.

Dimension	Examples	E-mail		Blog		Written	
Words > 6 letters	*solution, diversity*	12.69	(1)	15.30	(1)	13.06	(2)
Present tense	*meets, goes*	11.12	(2)	9.96	(2)	13.95	(1)
First-person sing.	*I, me*	6.51	(3)	6.81	(4)	10.63	(3)
Social processes	*talk, friend*	6.34	(4)	5.90	(5)	6.51	(4)
Inclusive	*and, with*	6.32	(5)	5.77	(6)	5.95	(5)
Articles	*a, the*	6.17	(6)	6.84	(3)	4.73	(6)
Past tense	*met, went*	4.56	(7)	4.06	(7)	3.79	(8)
Exclusive	*but, without*	3.55	(8)	3.61	(8)	4.21	(7)
Positive emotions	*happy, good*	3.10	(9)	2.86	(9)	3.38	(9)
Tentative	*maybe, perhaps*	2.62	(10)	2.43	(10)	2.84	(10)
Discrepancy	*could, would*	2.18	(11)	1.94	(11)	2.84	(11)
Negations	*no, never*	1.69	(12)	1.83	(12)	2.18	(13)
Insight	*think, know*	1.65	(13)	1.71	(13)	2.47	(12)
Negative emotions	*hate, worthless*	0.99	(14)	1.66	(14)	1.80	(14)
Causation	*because, hence*	0.68	(15)	0.73	(15)	1.10	(15)

Note: Ordered by rank of E-mail data. Written data reproduced from Pennebaker & King (1999:1302), Table 2.

singular pronouns, but few articles. One variable in the blog data, and two in the e-mail corpus, actually falls below the criterion of minimum 1% usage. To ensure compatibility with the factor analysis of Pennebaker & King, we adopt the same selection criteria. However, the fourth criterion (requiring at least 1% usage) is not met by: causation, in either the e-mail or blog corpora (.73% and .68%, respectively); and negative emotion, in the e-mail corpus (.99%).[2] For consistency of comparison between our e-mail and blog corpora, we take the more conservative measure and perform the analysis using 13 variables, omitting the causation *and* negative emotion categories.[3]

[2] Indeed in the blog data, although negative emotion words received 1.66% usage, this was the second lowest mean of the 15 variables, after causation.

[3] Exploratory factor analysis was also carried out using all 15 of Pennebaker & King's variables, regardless of their mean usage. This resulted in a 4-factor solution,

Table 2

Rotated factor loadings for exploratory analysis of LIWC variables - E-mail data.

	Factor 1:	Factor 2:	Factor 3:
Dictionary	(17.0% var)	(17.0% var)	(14.2% var)
Exclusive	.697		
Negations	.598		
Discrepancies	.593		
Tentative	.581		
Inclusive	−.561		−.457
Present tense		.812	
Articles		−.710	
1st-psn-sng		.622	
Words>6 lett		−.556	
Insight			
Past tense			.755
Social			.658
Pos emotions			.577

Note: Only loadings of .40 or above are shown. $N = 105$.

Exploratory factor analysis for the e-mail and blog data was carried out on the means of each subject's texts. Diagnostic tests reveal that present datsets have a similar suitability for this approach as in the original study.[4]

Results

The scree plot for these data indicated that a three factor model would best fit the data (in both cases four factors that had eigenvalues over 1). Principal-component analysis extracted the factors, and varimax rotation was used to en-

with all 15 variable having communalities greater than .37 for the e-mail data, and greater than .40 for the blog data. We note however that the factor analyses with 13 variables most closely resemble the original factors derived from written data, and thus we report this analysis in detail.

[4] Bartlett's test of sphericity reveals significant scores ($p < .001$) while Kaiser-Meyer-Olkin's measurement of sampling adequacy is greater than .5 in all instances.

Table 3

Rotated factor loadings for exploratory analysis of LIWC variables - Blog data.

Dictionary	Factor 1: (21.0% var)	Factor 2: (20.2% var)	Factor 3: (12.4% var)
Exclusive	.833		
Discrepancies	.762		
Tentative	.728		
Negations	.674		
Articles		−.830	
1st-psn-sng		.667	
Present tense	.488	.638	
Words>6 lett		−.608	
Insight		.444	
Pos emotions			.670
Social			.622
Inclusive			.600
Past tense			.453

Note : Only loadings of .40 or above are shown. $N = 71$.

able interpretation. For the e-mail data, all variables had communality greater than .35 with the exception of insight words (.26). All blog variables had communality greater than .33. The rotated factor loadings for e-mail and blog corpora are shown in Tables 2 and 3 respectively.

Discussion

In order to aid comprehension, Table 4 shows the direction of the factor loadings for the 13 variables of e-mail and blog data, and the loadings for 15 variables of the written data (Pennebaker & King's additional variables excluded from 13 variable analysis are indicated in italics). Here we find that the ordering of Factors 1 and 2 switches in the 13 variable e-mail and blog analyses in comparison to that in the written study.

Factor 1 of both the e-mail (eigenvalue = 2.22) and blog (eigenvalue = 2.74)

Table 4

Direction of loading of e-mail and blog data with 13, and in the original study using 15, LIWC variables.

	1e	1b	2w	2e	2b	1w	3e	3b	3w	4w
Exclusive	+	+	+							
Discrepancies	+	+	+			+				
Tentative	+	+	+							
Causation										+
Negations	+	+	+							
Articles				−	−	−				
1st-psn-sng				+	+	+				
Present tense		+		+	+	+			+	
Words>6 lett				−	−	−				
Insight					+					+
Pos emotions							+	+	−	
Social							+	+	+	
Inclusive	−		−				−	+		
Past tense							+	+	+	
Neg emotions										−

Note: e = E-mail; b = Blog; w = Written. Factors 1 and 2 of the written data (Pennebaker & King) are switched, compared with the e-mail and blog data. Italicised variables are those excluded from the current study. Written data reproduced from Pennebaker & King (1999:1303), Table 3.

data, includes positive loadings for negation and exclusive, discrepancy, and tentative words, with the e-mail data also showing negative loading for inclusive words, matching Factor 2 of the written data exactly. The blog data adds a positive loading for present tense words.

Factor 2 again is common to e-mail (eigenvalue = 2.21) and blog (eigenvalue = 2.63) data, with both showing loading of first-person singular and present tense, and a negative loading of articles and words longer than 6 letters. Here we find that insight loads onto Factor 2 in the blog data, but does not load strongly onto any e-mail data factors—although in the written data, Pennebaker & King found this loads onto their Factor 4. Also, discrepancies load on the equivalent factor in the written data.

Factor 3 of e-mail (eigenvalue = 1.85) and blog (eigenvalue = 1.61) analyses show loadings of positive emotions (although negative in the written study), social and past tense, although here e-mail data shows a negative loading of inclusive words, whereas this is a positive loading for blog data. Also in the written data there are loadings of present tense words, but none for inclusive words.

Examining the overall variance described by this analysis, we note that the three factors account for the most variance in the blog data, at 53.6%, compared with the e-mail data at 48.2%, and finally (discounting their fourth factor) Pennebaker & King's written data with 42.5% (51.1% if their fourth factor is included).

We therefore note that the three factors derived here closely match the first three factors of Pennebaker & King. Their 'Making Distinctions' factor is identical to the first factor derived from the e-mail data, and with one exception, the blog data also. For 'Immediacy' the only differences between the e-mail and blog data are that insight was not found to load for e-mail; in neither case—unlike the written data—did discrepancies load. Considering Pennebaker & King's 'The Social Past' factor, this is also very similar, but with a few exceptions: The past tense and social words are positively loaded across all studies, while positive emotion loads positively in both e-mail and blog data, rather than negatively as in the study of Pennebaker & King. Present tense words, which loaded on 'The Social Past' in the written study, are absent from both e-mail and blog analyses. By contrast, inclusive words did not load on this factor in the written study, but do in the CMC data: positively for blogs, and negatively for e-mail.

Experiment 2: CMC factor structure and personality traits

Following Pennebaker & King's procedure for investigating written language, this section studies correlations with personality. It uses the factors derived in the previous section, and the related variables, to help explore language differences within personality dimensions.

Analytic method

In the previous section, we replicated Pennebaker & King's factor analysis with causation and negative emotion words removed since they did not meet the mean usage criterion. Therefore we compare this 13 variable solution for e-mail and blog data with the original study. This also means that there will not be a comparison of the fourth factor in any detail, since it is the factor on which there is least agreement (Pennebaker, 2004). It should be noted that due to vastly different population sizes (e-mail data's 105 participants, and the 71 contributors to the blog corpus, contrast strongly with Pennebaker & King's 841 participants) it is difficult to compare correlation strengths. So aside from the factors, we focus upon the correlations that reach significance, as well as more general differences in direction of relationship.

Results

The correlation results of the 13 variable factor analyses with personality can be seen in Tables 5 and 6. Table 7 allows comparison between the signs of the correlations, and also those from Pennebaker & King's original study; results from the latter are more fully reproduced in Table 9.

Discussion

We perform comparisons for each personality trait for each factor, beginning with Neuroticism and Extraversion, since these were in all three studies, followed by Openness, Agreeableness and Conscientiousness, since these allow comparison with the original written study. We finally discuss the findings for Psychoticism, the third trait of the EPQ-R model, after Extraversion and Neuroticism.

Neuroticism

In the blog data, we find that Factor 1 ('Making Distinctions') correlates significantly with Neuroticism (p <.05), along with the constituent loading variable discrepancies (p <.01), reflecting the general positive trend. In contrast, the e-mail data represents a generally weak negative trend between Factor 1 and Neuroticism, with only the constituent loading variable inclusive words showing a strong (positive) correlation (p <.01; the strongest relationship found for the e-mail data). No notable relationships are present between the relevant factor 'Making Distinctions' in the written data and this personality trait.

Table 5

Correlation of LIWC factors (13 variables) with personality scores - e-mail data.

		EPQ-R Dimension	
	Neuroticism	Extraversion	Psychoticism
Factor 1	**−.11**	**−.03**	**.11**
Exclusive	−.02	−.10	−.01
Negations	−.03	−.08	−.02
Discrepancies	.04	.09	.13
Tentative	−.14	.00	.13
– Inclusive	.26**	−.02	−.11
Factor 2	**.12**	**−.08**	**−.11**
Present tense	.14	−.10	−.06
– Articles	−.02	.11	.12
1st-psn-sng	.16	−.12	−.23*
– Words>6 lett	.04	−.05	−.01
Factor 3	**−.24***	**.11**	**.04**
Past tense	−.19	.06	−.09
Social	−.05	.01	.02
Pos emotions	−.13	.15	.07
– *Inclusive*	*.26***	*−.02*	*−.11*

Note: $N = 105$. Italics are used to indicate variables loading on a second factor. '–' is used to indicate a negative factor loading. LIWC categories are ordered as they load onto their Factor. *$p < .05$. **$p < .01$, two tailed.

Factor 2, or 'Immediacy' shows a generally positive relationship with Neuroticism in both e-mail and written studies (although this only reaches levels of significance in the latter study, for the factor itself and first person singular words, and—negatively—articles). No notable relationships exist in the blog data.

For Factor 3 ('The Social Past'), there is a significant negative correlation with Neuroticism in the e-mail data (p <.05), with other variables loading on this factor also generally showing a negative trend; the same is true for the blog data, however no relationships reach significance. For the written data,

Table 6

Correlation of LIWC factors (13 variables) with personality scores - blog data.

		Five-factor Dimension				
		N	E	O	A	C
Factor 1		**.24***	**–.19**	**–.06**	**–.25***	**.02**
	Exclusive	.13	–.08	.14	–.19	–.06
	Discrepancies	.34**	–.25*	–.12	–.29*	–.04
	Tentative	.14	–.14	–.11	–.20	.06
	Negations	.16	.02	–.22	–.24*	.10
	Present tense	.16	.20	.01	–.09	.10
Factor 2		**.00**	**.17**	**–.18**	**–.14**	**.07**
–	Articles	–.07	.03	.14	.26*	–.05
	1st-psn-sng	–.02	.18	–.10	–.08	–.06
	Present tense	*.16*	*.20*	*.01*	*–.09*	*.10*
–	Words>6 lett	.02	–.06	.29*	.26*	.03
	Insight	–.11	.06	.11	.09	.08
Factor 3		**–.08**	**.18**	**.30***	**.13**	**–.15**
	Pos emotions	–.04	.16	.13	.07	–.06
	Social	–.04	.24*	.20	.04	–.11
	Inclusive	–.01	.02	.25*	.09	–.09
	Past tense	.01	–.12	–.03	–.12	–.16

Note: N = 71, two tailed, *p<0.05, **p<0.01. Italics are used to indicate variables loading on a second factor. '–' is used to indicate a negative factor loading.

although there is little overall relationship with Neuroticism, positive emotion words which load onto the factor show a strong negative relationship, similar to that of the e-mail data (although as noted above, not significant). Although not included in the present analysis, we additionally note the strong positive correlation with negative emotion words in the written data (loading on Factor 4, 'Rationalization').

Table 7

Comparison of the correlation of LIWC factors with personality scores - across e-mail, blog and written data.

		Personality Dimension												
		N			E			O		A		C		P
		e	b	w	e	b	w	b	w	b	w	b	w	e
Making Dist.		−	+	+	−	−	−	−	+	−	−	+	−	+
	Exclusive	−	+	+	−	−	−	+	−	−	−	−		−
	Discrepancies	+	+	+	+	−	−	−	−	−	−	−	−	+
	Tentative	−	+	−	−	−	−	−	+	−	−	+	+	+
	Negations	−	+		−	+	−	−	−	−	−	+	+	−
	Present tense	+	+	−	+	+	+	+	−	−	+	+		−
Immediacy		+		+	−	+	+	−	−	−	+	+	−	−
−	Articles	−	−	−	+	+	−	+	+	+	−	−	−	+
	1st-psn-sng	+	−	+	−	+	+	−	−	−	+	−	+	−
−	Words>6 lett	+	+	−	−	−	−	+	+	+	−	+	+	−
	Insight		−	+		+	−	+	+	+	+	+	−	
The Social Past		−	−	+	+	+		+	+	+	−	−	−	+
$(-^{w})$	Pos emotions	−	−	−	+	+	+	+	−	+	+	−	+	+
	Social	−	−	−	+	+	+	+	+	+		−	+	+
$(-^{e,w})$	Inclusive	−	−	+	−	+	+	+	+	+	+	−	−	−
	Past tense	−	+	+	+	−	+	−	−	−	+	−	−	−

Note: Variables are placed under the factors emerging from the blog study. Negative loadings revealed across other studies are indicated within parentheses, with superscript initials indicating the relevant study. Factors are named according to Pennebaker & King (1999).

Extraversion

With the exception of discrepancies in the blog corpus (p <.05), Factor 1 ('Making Distinctions') and loading variables do not show more than a negative trend in relationship to Extraversion (with this stronger for the blog data). In contrast, the written data of Pennebaker & King shows a much stronger significant negative correlation (with the exception of inclusive words, which correlate positively).

Factor 2 ('Immediacy') showed very few consistent or significant relationships with Extraversion across the three studies (blogs being positively related, writing slightly positive, and e-mails negative), with the exception being the variable articles in the written study, which showed a negative relationship.

Although for Factor 3 ('The Social Past') both e-mail and blog data show a general positive trend, this is not the case with the written study, although social and positive emotion constituent variables both show a strong relationship with Extraversion. In the case of the blog copus, only social words related significantly positively (p <.05).

Openness

Turning now to the remaining five-factor personality variables, we will solely compare the findings of the blog and written data. In both of these studies Factor 1, or 'Making Distinctions', is not significantly correlated with Openness, although in the written data both exclusive and negation variables correlate negatively.

For Factor 2 ('Immediacy'), the blog data shows a negative trend with Openness; indeed, the most notable relationship is a positive one, with a negatively loaded variable, words of more than 6 letters (p <.05). In contrast, this factor shows the strongest of all relationships in the written data of Pennebaker & King, where it is negatively related to Openness, along with first-person singular and present tense words, while it is positively related to articles and words of more than 6 letters, which load negatively on the Immediacy factor.

Both studies find a significant positive correlation with Factor 3, 'The Social Past' and Openness, and in fact for the blog data, this represents the strongest factor correlation (r=.295; p <.05). The blog study also shows a significant positive relatioinship between Openness and the inclusiveness variable (p <.05).

Agreeableness

Factor 1, ('Making Distinctions') and Agreeableness demonstrate another strong relationship in the blog data, which in this case is negative (p <.05), with constituent variables discrepancies and negations also similarly showing a negative relationship (p <.05). In the written data, the factor is not significantly correlated, a number of the constituent variables are, and the overall pattern is very similar to that in the blog data.

Factor 2 ('Immediacy') in the written data shows a positive relationship with Agreeableness, along with constituent variable first-person singular, and a negative relation to articles. Although the factor does not correlate significantly in the blog study, it actually shows a strong trend towards a negative relationship with this trait, and appears to demonstrate the largest area of disagreement between the two studies. The constituent variables articles and words of more than 6 letters both correlate positively and significantly with Agreeableness (p <.05).

Factor 3 ('The Social Past') shows a very small negative relationship with Agreeableness in the written data, whereas the blog data apparently shows a larger (but still non-significant) positive correlation. Pennebaker & King do, however, note a positive relationship with positive emotion words.

Conscientiousness

Although the blog data does not appear to demonstrate much interaction between Factor 1 ('Making Distinctions') and Conscientiousness, in the written data, there is a strong negative relationship at the factor level.

Neither Factor 2 ('Immediacy') of the blog or written corpus bears much relationship with Conscientiousness.

Factor 3 ('The Social Past') shows a weak non-significant negative relationship with Conscientiousness in both blog and written data; in the case of the latter, the negatively loaded constituent variable positive emotion words correlates positively with this trait.

Psychoticism

In order to relate the findings for Psychoticism to those of the five-factor model used in the blog corpus, and also by Pennebaker & King (1999), we take Psychoticism to be inversely related to Agreeableness and Conscientiousness. For Factor 1 ('Making Distinctions'), we find that Psychoticism shows a trend towards a positive relationship with this factor, which is consistent (when inverted) with the negative relationships found in the written data across Agreeableness and Conscientiousness, and in the case of the blog data, consistent with Agreeableness.

For Factor 2 ('Immediacy'), Psychoticism shows a non-significant negative relationship with this factor, and also relates significantly negatively to the first-person singular variable ($p < .05$), with this finding consistent (when inverted) with the significant positive relationship found with Agreeableness in the written data. However, we note that this is less consistent with the non-significant negative relationship found in the blog data.

Factor 3 (‘The Social Past’) of the e-mail data shows little relationship to Psychoticism, and although we cannot infer too much from this, we note that this is generally consistent with the blog and written data findings for Agreeableness and Conscientiousness.

General discussion

In this study, we addressed two sets of research questions (Section 3). We now address these in turn, drawing together the material from each experiment.

E-mail versus Blog factors; CMC versus Non-CMC

The primary aim of this study has been to investigate how the different CMC genres of e-mail and blog relate, and to contrast them with non-CMC data from a previous study. What are very apparent, firstly, are the broad similarities in factor structure between these three language varieties. Although there has been much speculation about the divergence in language varieties brought about by the advent of CMC, it should not be surprising that these factors, originally derived from writing, have been replicated here. Rather, it is a tribute to Pennebaker & King, the authors of the original study, for their effective selection of LIWC variables which have indeed demonstrated topic independence and apparent robustness across different communication genres. We do note, however, that consistent with Pennebaker’s own findings, the first two factors are replicated more closely than the third, and fourth factors

(Pennebaker, 2004).

This is not to say that there are *no* differences between these genres, as we previously note in our discussion of the results (Sections 5.3 and 6.3). Examination of directions of variable loadings upon the factors across the three studies shows some minor variation: From this, it is interesting to note that e-mail appears to occupy the middle ground between the blog and written genres: There is no case where variable loadings are shared by blog and written language factors, but not by e-mail—yet there are cases where e-mail and one of the other genres together differ from the third genre. Here we note that e-mail and written language are more similar in their use of inclusive words (like *and* and *with*), which loads on the 'Making Distinctions' factor, but that e-mail and blog language oppose the written finding for 'The Social Past' factor because they have *positive*, rather than *negative* loadings for positive emotion words.

There are also distinguishing features in the way variables load onto the factors: Characteristic features of weblog language are the positive loading of present tense words on the 'Making Distinctions' factor, and insight onto the 'Immediacy' factor; conversely, written language is characterised by the positive loadings of discrepancies on the 'Immediacy' factor and by present tense on 'The Social Past' factor. All three genres are apparently distinguished by inclusive words in relation to 'The Social Past' with blogs showing positive loading, e-mails showing negative loading, and written language an absence of loading for this factor.

In summary, we therefore observe there are broad similarities between these genres with differences in factor structure which distinguish between all three, although in terms of factor structure, e-mail appears intermediate between the blog and written texts. Like Pennebaker & King, we selected the variables in the current study in order to ensure comparability across genres. However, we also note that in order to identify linguistic features which distinguish these different genres, future research may focus on LIWC variables which were excluded from the original study for their lack of consistency across genres. An alternative approach which would highlight characteristic features is the data-driven technique which has already successfully been applied to e-mail and blog data (Gill, 2004; Nowson, 2006; Oberlander and Gill, 2006).

Personality and factors in CMC; and five versus three factors

The second aim of this study was to investigate the features of personality across the three genres, in particularly in terms of CMC in comparison to non-CMC language. Primarily we focus on the traits of Extraversion and Neuroticism, since these are common to the three studies.

Taking the findings from e-mail and blogs together, we note that Neuroticism shows a positive relationship with 'Making Distinctions', and a negative relationship with 'The Social Past', whereas in the non-CMC written data, it correlates positively with the 'Immediacy' factor. For Extraversion, although the CMC data does not show significant relationship to any of the factors, it does show a tendency in the same direction as the negative relationship with 'Making Distinctions' found in the non-CMC data.

Turning now to the other factors, we note that Openness reveals agreement between blog CMC and non-CMC data in terms of a positive relationship with

‘The Social Past’, with the most significant non-CMC (negative) relationship with ‘Immediacy’ also mirrored (non-significantly) in the blog CMC data. For Agreeableness, the negative relationship with ‘Making Distinctions’ in blog CMC data contrasts with the positive relationship with ‘Immediacy’ in non-CMC data; the negative relationship between Conscientiousness and ‘Making Distinctions’ in the non-CMC study is not found here.

Although we expected that additional stronger relationships would emerge with the inclusion of Psychoticism, this was not the case, with no factors reaching significance in correlation. Psychoticism was generally compatible (when inverted) with the other studies for ‘Making Distinctions’ and ‘Immediacy’, and like the other studies showed little relationship to ‘The Social Past’.

We therefore note that although there are general consistencies between CMC and non-CMC data when they are related to personality measures, conclusive comparison is difficult due to fewer of the relationships reaching significance in the CMC data, apparently as a result of having many fewer participants in these studies (105 and 71 for the e-mail and blog data, respectively, compared to 841 in Pennebaker & King’s written study). Of the significant relationships observed, we note the different behaviour in the CMC data leading to a positive relationship between Neuroticism and ‘Making Distinctions’ and a negative one between it and ‘The Social Past’, compared with non-CMC’s positive relationship with ‘Immediacy’. For Agreeableness, CMC relates negatively to ‘Making Distinctions’, rather than positively to ‘Immediacy’.

We can speculate as to the reasons for these differences. First, it is rather hard to see why ‘The Social Past’ should vary more in CMC, but perhaps the differences are associated with the differing role which positive emotions play in this factor in CMC, as opposed to non-CMC. Secondly, however, 'Immediacy' seems to vary less in CMC, and this may be attributable to an audience design effect, particularly with blogs. Because both e-mail and blogs are written for potentially remote audiences, all authors rely less on linguistic devices involving immediacy. With less use of them overall, there is less variation to be found amongst individual authors. Secondly, it seems that 'Making Distinctions' varies more in CMC, and the reason may be related to reduced social presence. According to Nowson (2006), bloggers often state that they write in order to 'vent'; equally, e-mail is often held to encourage the freer expression of critical opinions. If this is so, then CMC authors may be more likely to indulge in critical writing. With more criticism going on, there will be more variation to be found among individual authors. It seems plausible that the more anxious and less agreeable authors will be those most likely to draw attention to faults and discrepancies.

Conclusion

In this study we have replicated a previously derived factor structure for language in written (non-CMC) environments in two types of CMC: e-mail and blogs. The factors are largely similar, this being a merit of the original analytical design. However we note the value of particular features for distinguishing between CMC and non-CMC communication. Furthermore, at a more detailed level, we observe that the patterning of such variation across genres indicates that e-mail shares the most similarity with both blog and written language, and as such mediates in style between the two, with blog and non-CMC written language more readily distinguished. We note the value of selecting different variables, or data-driven analyses, to further highlight distinguishing characteristics of these genres.

When language factors are correlated with personality measures, we find different patterns of language behaviour for Neuroticism and Agreeableness in CMC versus non-CMC environments. Replacing Openness, Agreeableness and Conscientiousness of the five-factor model with Psychoticism of the EPQ-R does not uncover significant new behaviour, or uncover stronger relationships, although findings are generally consistent with the five-factor approach. We do however note that the smaller number of participants in the CMC studies, compared with those in the previous non-CMC written study, results in fewer relationships reaching significance.

Acknowledgments

We thank Jeff Hancock and Matthias Mehl for very helpful comments on drafts of this paper. The first and second authors gratefully acknowledge studentship support from the UK Economic and Social Research Council. The first author is currently supported by a Faber Post-Doctoral Fellowship (05 512 AA 06 S2469).

References

Bälter, O. (1998). *Electronic Mail in a Working Context*. Ph.D. thesis, Royal Institute of Technology, Stockholm.

Baron, N. (1998). Letters by phone or speech by other means: the linguistics of email. *Language and Communication*, **18**, 133–170.

Baron, N. (2001). Commas and canaries: the role of punctuation in speech and writing. *Language Sciences*, **23**, 15–67.

Bell, A. (1984). Language as audience design. *Language in Society*, **13**, 145–204.

Belo, R. (2004). Blogs take on the mainstream. Available from BBC News online at http://news.bbc.co.uk/1/hi/technology/4086337.stm. Accessed June 9, 2006.

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

Biber, D. (1995). *Dimensions of Register Variation*. Cambridge University Press, Cambridge.

Biber, D. (2004). Towards a typology of web registers: a multi-dimensional analysis. Invited lecture, Conference of Corpus Linguistics: Perspectives for the future. Heidelberg University.

Bloch, J. (2002). Student/teacher interaction via email: the social context of internet discourse. *Journal of Second Language Writing*, **11**, 117–134.

Block, J. (1995). A contrarian view of the five-factor approach to personality description. *Psychological Bulletin*, **117**, 187–215.

Brown, G. and Yule, G. (1983). *Discourse analysis*. Cambridge University Press, Cambridge.

Buchanan, T. (2001). Online implementation of an IPIP five factor personality inventory [web page]. http://users.wmin.ac.uk/∼buchant/wwwffi/introduction.html [Accessed 25/10/05].

Buss, A. and Finn, S. (1987). Classification of personality traits. *Personality and Social Psychology*, **52**, 432–444.

Carment, D. W., Miles, C. G., and Cervin, V. B. (1965). Persuasiveness and persuasibility as related to Intelligence and Extraversion. *British Journal* *of Social and Clinical Psychology*, **4**, 1–7.

Cassell, J. and Tversky, D. (2005). The language of online intercultural community formation. *Journal of Computer-Mediated Communication*, **10**(2), article 2.

Cho, N. (1996). Linguistic features of electronic mail: Results from a pilot study. Paper presented at the Australia and New Zealand Communication Association Annual Conference, Brisbane.

Cohn, M., Mehl, M., and Pennbaker, J. (2004). Linguistic markers of psychological change surrounding september 11. *Psychological Science*, **15**, 687–693.

Colley, A. and Todd, Z. (2002). Gender-linked differences in the style and content of e-mails to friends. *Journal of Language and Social Psychology*, **21**, 380–392.

Collot, M. and Belmore, N. (1996). Electronic language: A new variety of English. In S. Herring, editor, *Computer-mediated communication: Linguistic, social and cross-cultural perspectives*, pages 13–28. Benjamins, Amsterdam.

Costa, P. and McCrae, R. R. (1992). *NEO PI-R Professional Manual*. Psychological Assessment Resources, Odessa, Florida.

Crowston, K. and Williams, M. (2000). Reproduced and emergent genres of communication on the world wide web. *The Information Society*, **16**(3), 201–216.

Deary, I. and Matthews, G. (1993). Personality traits are alive and well. *The Psychologist*, **6**, 299–311.

Dewaele, J.-M. (2001). Interpreting the maxim of quantity: interindividual and situational variation in discourse styles of non-native speakers. In E. Nèmeth, editor, *Selected Papers from the 7th International Pragmatics Conference*, volume 1, pages 85–99. International Pragmatics Association, Antwerp.

Dewaele, J.-M. and Furnham, A. (1999). Extraversion: The unloved variable in applied linguistic research. *Language Learning*, **49**, 509–544.

Dewaele, J.-M. and Furnham, A. (2000). Personality and speech production: a pilot study of second language learners. *Personality and Individual Differences*, **28**, 355–365.

Digman, J. (1990). Personality structure: Emergence of the five-factor model. *Annual Review of Psychology*, **41**, 417–440.

Efimova, L. and de Moor, A. (2005). Beyond personal webpublishing: An exploratory study of conversational blogging practises. In *Proceedings of the 37th Annual HICSS Conference*, Big Island, Hawaii.

Eysenck, H. (1970). *The Biological Basis of Personality*. Thomas, Springfield, IL.

Eysenck, H. (1993). From DNA to social behaviour: conditions for a paradigm of personality research. In J. Hettema and I. Deary, editors, *Foundations of personality*. Kluwer, Dordrect.

Eysenck, H. and Eysenck, S. B. G. (1991). *The Eysenck Personality Questionnaire-Revised*. Hodder and Stoughton, Sevenoaks.

Eysenck, S., Eysenck, H., and Barrett, P. (1985). A revised version of the psychoticism scale. *Personality and Individual Differences*, **6**, 21–29.

Fallows, D. (2005). How women and men use the internet. Technical report, Pew Internet and American Life Project.

Funder, D. (2001). Personality. *Annual Review of Psychology*, **52**, 197–221.

Furnham, A. (1990). Language and personality. In H. Giles and W. Robinson, editors, *Handbook of Language and Social Psychology*, pages 73–95. Wiley, Chichester.

Gifford, R. and Hine, D. W. (1994). The role of verbal behaviour in the encoding and decoding of interpersonal dispositions. *Journal of Research in Personality*, **28**, 115–132.

Gill, A. (2004). *Personality and Language: The projection and perception of personality in computer-mediated communication*. Ph.D. thesis, University of Edinburgh.

Gill, A. and Oberlander, J. (2002). Taking care of the linguistic features of extraversion. In *Proceedings of the 24th Annual Conference of the Cognitive Science Society*, pages 363–368.

Gill, A. J., Oberlander, J., and Austin, E. (2006). Rating e-mail personality at zero acquaintance. *Personality and Individual Differences*, **40**, 497–507.

Goldberg, L. R. (1993). The structure of phenotypic personality traits. *American Psychologist*, **48**(1), 26–34.

Gottschalk, L. A. and Gleser, G. C. (1969). *The measurement of psychological states through the content analysis of verbal behavior*. University of California Press, Berkeley.

Graybeal, A., Sexton, J., and Pennebaker, J. (2002). The role of story-making in disclosure writing: The psychometrics of narrative. *Psychology and Health*, **17**, 571–581.

Groom, C. and Pennebaker, J. (2005). The language of love: Sex, sexual orientation, and language use in online personal advertisements. *Sex Roles*, **52**, 447–461.

Gruber, H. (2000). Scholarly email discussion list postings: a single new genre of academic communication? In L. Pemberton and S. Shurville, editors, *Words on the Web*, pages 36–43. Intellect Books, Exeter.

Hancock, J. and Dunham, P. (2001a). Impression formation in computer-mediated communication. *Communication Research*, **28**, 325–347.

Hancock, J. and Dunham, P. (2001b). Language use in computer-mediated communication: The role of coordination devices. *Discourse Processes*, **31**, 91–110.

Hancock, J., Curry, L., Goorha, S., and Woodworth, M. (2006). On lying and being lied to: An automated linguistic analysis of deception. *Discourse Processes*, **in press**, xxx–yyy.

Henning, J. (2005). The blogging geyser. Technical report, Perseus.

Huffaker, D. (2004). *Gender similarities and differences in online identity and language use among teenage bloggers*. Master's thesis, Graduate School of Arts and Sciences, Georgetown University.

Joyce, E. and Kraut, R. (2006). Predicting continued participation in newsgroups. *Journal of Computer-Mediated Communication*, **11**(3), article 3.

Kilgarriff, A. (2001). Comparing corpora. *International Journal of Corpus Linguistics*, **6**, 231–245.

Kline, P. (1993). *The Handbook of Psychological Testing*. Routledge, London.

Krishnamurthy, S. (2002). The multidimensionality of blog conversations: The virtual enactment of September 11. Paper presented at Internet Research 3.0, Maastricht, The Netherlands.

Lippa, R. and Dietz, K. (2000). The relations of gender, personality, and intelligence to judges' accuracy in judging strangers' personality from brief video segments. *Journal of Nonverbal Behavior*, **24**, 25–43.

Lowe, W. (2004). Content analysis and its place in the (methodological) scheme of things. *Quantitative Methods*, **2**, 25–27.

Madden, M. (2006). Internet penetration and impact. Technical report, Pew Internet and American Life Project.

Marlow, C. (2004). Audience, structure and authority in the weblog community. Presented at The International Communications Association Conference, New Orleans.

Matthews, G., Deary, I. J., and Whiteman, M. C. (2003). *Personality Traits*. Cambridge University Press, Cambridge, 2nd edition.

McCrae, R. and Costa, P. (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology*, **52**, 81–90.

McCrae, R. and Costa, P. (1997). Personality trait structure as a human universal. *American Psychologist*, **52**, 509–516.

Mehl, M. (2005). Quantitative text analysis. In M. Eid and E. Diener, editors, *Handbook of multimethod measurement in psychology*, pages 141–156. American Psychological Association, Washington, DC.

Mehl, M. and Pennebaker, J. (2003). The sounds of social life: A psychometric analysis of student's daily social interactions. *Journal of Personality and Social Psychology*, **84**, 857–870.

Mehl, M., Gosling, S., and Pennebaker, J. (2006). Personality in its natural habitat: Manifestations and implicit folk theories of personality in daily life. *Journal of Personality and Social Psychology*, **in press**, xxx–yyy.

Newman, M., Pennebaker, J., Berry, D., and Richards, J. (in press). Lying words: Predicting deception from linguistic styles. *Personality and Social Psychology Bulletin*, **29**, 665–675.

Nilsson, S. (2003). The function of language to facilitate and maintain social networks in research weblogs.

Nowson, S. (2006). *The Language of Weblogs: A study of genre and individual differences*. Ph.D. thesis, University of Edinburgh.

Nowson, S., Oberlander, J., and Gill, A. J. (2005). Weblogs, genres and individual differences. In *Proceedings of the 27th Annual Conference of the Cognitive Science Society*, pages 1666–1671.

Oberlander, J. and Gill, A. J. (2006). Language with character: A stratified corpus comparison of individual differences in e-mail communication. *Discourse Processes*, **42**, 239–270.

Oxman, T., Rosenberg, S., Schnurr, P., and Tucker, G. (1988). Diagnostic classification through content analysis of patients' speech. *American Journal of Psychiatry*, **145**, 464–468.

Panteli, N. (2002). Richness, power cues and email text. *Information & Management*, **40**, 75–86.

Pennebaker, J. (2004). Personal correspondence.

Pennebaker, J., Mehl, M., and Niederhoffer, K. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology*, **54**, 547–577.

Pennebaker, J. W. (1997). Writing about emotional experiences as a therapeutic process. *Psychological Science*, **8**, 162–166.

Pennebaker, J. W. and Francis, M. (1999). *Linguistic Inquiry and Word Count (LIWC)*. Lawrence Erlbaum Associates, Mahwah, NJ.

Pennebaker, J. W. and King, L. (1999). Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology*, **77**, 1296–1312.

Pennebaker, J. W., Mayne, T., and Francis, M. (1997). Linguistic predictors of adaptive bereavement. *Journal of Personality and Social Psychology*, **72**, 863–871.

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

Pytlik Zillig, L., Hemmenover, S., and Dienstbier, R. (2002). What do we assess when we assess a Big 5 trait? a content analysis of the affective, behavioural, and cognitive processes represented in Big 5 personality inventories. *Personality and Social Psychology Bulletin*, **28**, 847–858.

Rayson, P. (2003). *Wmatrix: A statistical method and software tool for linguistic analysis through corpus comparison*. Ph.D. thesis, Lancaster University.

Santini, M. (2005). Clustering web pages to identify emerging textual patterns. RECITAL 2005, Dourdan.

Scherer, K. (1979). Personality markers in speech. In K. R. Scherer and H. Giles, editors, *Social Markers in Speech*, pages 147–209. Cambridge University Press, Cambridge.

Shepherd, M., Watters, C., and Kennedy, A. (2004). Cybergenre: Automatic identification of home pages on the web. *Journal of Web Engineering*, **3**(3&4), 236–251.

Smith, C. (1992). Introduction: inferences from verbal material. In C. Smith, editor, *Motivation and personality: Handbook of thematic content analysis*, pages 1–17. Cambridge University Press, Cambridge.

Thorne, A. (1987). The press of personality: A study of conversations between introverts and extraverts. *Journal of Personality and Social Psychology*, **53**, 718–726.

Vazire, S. and Gosling, S. D. (2004). e-perceptions: Personality impressions based on personal websites. *Journal of Personality and Social Psychology*, **87**, 123–132.

Werry, C. (1996). Linguistic and interactional features of internet relay chat. In S. Herring, editor, *Computer-mediated communication: Linguistic, social and cross-cultural perspectives*, pages 47–63. Benjamins, Amsterdam.

Whittaker, S. (2003). Theories and methods in mediated communication. In A. Graesser, M. Gernsbacher, and S. Goldman, editors, *The Handbook of Discourse Processes*, pages 243–286. Lawrence Erlbaum Associates, Mahwah, New Jersey.

Wiggins, J. and Pincus, A. (1992). Personality: Structure and assessment.

Annual Review of Psychology, **43**, 473–504.

Wintertree (2000). *Sentry Spelling Checker Engine*. Wintertree Software, Ontario, Canada.

Yates, S. (1996). Oral and written linguistic aspects of computer conferencing: A corpus based study. In S. Herring, editor, *Computer Mediated Communication: Linguistic, social and cross-cultural perspectives*, pages 29–46. JohnBenjamin, Amsterdam.

Yellen, R., Winniford, M., and Sanford, C. (1995). Extraversion and introversion in electronically-supported meetings. *Information & Management*, **28**, 63–74.

Reproduced Tables Follow

Table 8. Rotated Factor Loadings for Exploratory Analysis of LIWC Dictionaries

	Factor 1:	Factor 2:	Factor 3:	Factor 4:
	Immediacy	Making Distinctions	The Social Past	Rationalization
Dictionary	(22.4% variance)	(10.3% variance)	(9.8% variance)	(8.6% variance)
First-person Sing.	.823			
Articles	−.765			
Words > 6 letters	−.683			
Discrepancies	.485	.427		
Exclusive		.674		
Tentative		.644		
Negations		.579		
Inclusive		−.463		
Past tense			.856	
Present tense	.593		.596	
Positive emotion			−.469	
Social			.425	
Insight			.627	
Causation				.598
Negative emotion				−.443

Note. Only loadings of .20 or above are shown. $N = 838$. (reproduced from Pennebaker and King, 1999, p. 1303).

Table 9

LIWC Factors and Simple Correlations with Five-Factor Scores

	Five-Factor Dimension				
LIWC factor	N	E	O	A	C
Immediacy	**.10***	**.04**	**−.16****	**.07***	**−.02**
First-person Sing.	.13**	.04	−.13**	.07**	.01
Articles	−.09*	−.09*	.13**	−.15**	−.04
Words > 6 letters	−.03	−.04	.16**	−.03	.06
Present tense	−.06	.01	−.15**	.04	.00
Discrepancies	.05	−.03	−.01	−.02	−.07*
Making Dist.	**.05**	**−.14****	**.06**	**−.05**	**−.13****
Exclusive	.10*	−.06	−.08*	−.08*	.00
Tentative	.11**	−.02	−.06	−.14**	.06
Negations	.00	−.04	−.15**	−.12**	.05
Inclusive	.01	.03	.06	.07*	−.01
The Social Past	**.04**	**.00**	**.08***	**−.02**	**−.04**
Past tense	.03	.04	−.03	.06	−.06
Social	−.01	.12**	.02	.00	.02
Positive emotion	−.13**	.15**	−.06	.07*	.07*
Rationalisation	**−.06**	**.02**	**−.03**	**.07**	**.04**
Insight	.03	−.02	.07*	.05	−.01
Causation	.03	−.08*	−.08*	.00	−.07*
Negative emotion	.16**	−.08*	.05	−.07*	−.15**

Note. $N = 841$. Two variables are coded onto two factors: Present tense is also part of The Social Past; Discrepancy is a part of Making Distinctions. The following variables are negatively loaded on their respective factors: Articles, Words of more than 6 letters, Inclusive, Present tense (for The Social Past only), and negative emotion.

$*p < .05$, $**p < .01$, two tailed.

(reproduced from Pennebaker and King, 1999, p. 1307).