

What Are They Blogging About? Personality, Topic and Motivation in Blogs

Alastair J. Gill

Center for Technology and Social Behavior
Northwestern University
2240 Campus Drive
Evanston, IL 60208, USA
alastair@northwestern.edu

Scott Nowson

Appen Pty Ltd,
Chatswood
Sydney, NSW 2067, Australia.
snowson@appen.com.au

Jon Oberlander

School of Informatics
University of Edinburgh
10 Crichton Street
Edinburgh, EH8 9AB, UK
J.Oberlander@ed.ac.uk

Abstract

Personal weblogs (blogs), provide individuals with the opportunity to write freely and express themselves online in the presence of others. In such situations, what do bloggers write about, and what are their motivations for blogging? Using a large blog corpus annotated with the LIWC text analysis program, we examine the content of blogs to provide insight into the role of personality in motivation for blogging. As predicted, we find that highly Neurotic authors use blogs to serve a cathartic or auto-therapeutic function, and reflect mainly upon themselves and negative emotions. Highly Extraverted blog authors, as expected, use blogs to document their lives at a high level, and uniquely interact directly with the reader. Additionally Extraverts use blogs to vent both positive and negative emotions. Bloggers who are high scorers on the Openness trait are more concerned with leisure activities, although they are more evaluative than intellectual, whereas high Conscientiousness bloggers tend to report daily life – and work – around them. As in other contexts, expressing positive rather than negative emotions is associated with high Agreeableness, but that trait is associated with self reference to a greater degree in blogs than elsewhere. In general, findings are consistent with other contexts indicating that bloggers tend to adapt to the possibilities of the medium, rather than try to present themselves differently.

Introduction

Personal weblogs (blogs), arguably more than any other media, have enabled ordinary people around the world to indulge their passion for writing and self expression. Individuals have always had the option of keeping a personal diary or journal in which to write their most intimate thoughts and feelings, with these rarely seeing the light of day and being guarded with some care. In rare cases, politicians or cultural commentators may have their diaries published, giving rise to suspicions that the diaries were written with eventual publication in mind (e.g., Samuel Pepys). Nowadays weblogs provide the opportunity for immediate publication (Sorapure, 2003).

A weblog is defined as a “frequently updated website consisting of dated entries arranged in reverse chronological order so the most recent post appears first” (Herman, Jahn and Ryan 2005). The most common form of weblog is the personal diary or journal blog (Herring, Scheidt, Sabrina Bonus, and Wright. 2005). Indeed the term blog is often seen as being synonymous with the online journal form (McNeill, 2005). Writing a blog – like a diary – provides the author with a great deal of freedom for personal expression: for example, what to write about, and also how objective or opinionated to be (Nowson, forthcoming). Given such freedom, blogs are an ideal medium for providing an insight into the most fundamental characteristics of the blog authors, namely their personalities. Therefore in this paper, we further investigate how personality influences language use in blogs, in particular examining how linguistic features reflect what individuals choose to write about, and how they present themselves (Goffman, 1959). We use such choices of presentation and topic to examine author motivation for blogging.

Background to Blogs

Blogs tend to be an informal genre with few expectations as to content or style. This means the language used within can be seen in its natural habitat, less constrained by rules or boundaries (Nowson, forthcoming). Like other forms of computer-mediated communication (CMC), blogs bear similarity to both spoken and written forms of communication (Gruber 2000; Yates 1996). However, weblogs are generally more polished, more grammatical, and better structured than, say, message boards and are therefore more closely related to publishing (Glance, Hurst and Tomokiyo, 2004). However, given the opportunities for readers to respond and leave comments, weblogs can be viewed as both monologue and dialogue (Efimova and de Moor 2005), and can therefore be thought of as both ‘simultaneously self-reflective thoughts presented publicly, and continuous conversations’ (Nilsson 2003, p. 31).

Examination of why bloggers are motivated to write a weblog reveal many drivers (Li 2005; Nardi, Schiano, Grumbrecht, and Swartz, 2004). Two principal reasons focus on the self: firstly, the desire of bloggers to

document their life; secondly, the use of writing as catharsis, whereby the authors vent and work through their emotions (Nardi et al. 2004). Another motivation for the author is the desire to discuss external topics of interest to them, expressing with varying degrees of objectivity, their opinions and feelings ('filters', Herring et al. 2005). In other cases motivation for blogging has been related back to psychological profiles of the author: exhibitionist, extrovert, creativity, auto-therapeutic, self-presentation, and self-promotion were found, based on an analysis of Polish blogs (Cywińska-Milonas, 2003, cited in Trammell, Tarkowski, Hofmokl, and Sapp, 2006). For similar promotion and presentation reasons businesses have started their own 'corporate weblogs' - in order to better connect with customers (Kelleher and Miller 2006; Puschmann, forthcoming).

Analysis of weblogs has cast light upon the relative influence of such motivations: Applying content analysis to Polish blogs revealed that self expression is a greater motivation than social interaction, and content is more likely to be internal feelings or thoughts than external hobbies or interests (Trammell et al. 2006). In other cases content analysis is used to determine the key topics of discussion (Glance et al. 2004; Nallapati and Cohen 2008), or mood of the author (Mishne, 2005) with more advanced techniques used to track discussion across the blogosphere. Additionally, studies have focused on *how* topics are discussed, for example, changes in language surrounding the events of September 11 2001 (Cohn, Mehl and Pennebaker, 2004). In this case, it was found that in the short term there were rapid increases in the expression of negative emotion, authors appeared to be more cognitively and socially engaged and wrote with greater psychological distance. Over time, these features slowly returned to their baseline levels.

Nowson, Oberlander and Gill (2005) used a grammatically-based measure of contextuality (Heylighen and Dewaele 2002) to compare personal weblogs to other genres from the British National Corpus (BNC). Personal weblogs were found to be less contextual than personal letters and mailing list postings but more contextual than biographical and non-academic writings in social science. This echoes findings reported earlier about the nature of blogging language (at least that in personal blogs) being somewhere between written and spoken.

Examining language use in a community of researchers' personal weblogs, Nilsson (2003) found a much higher use of *in-group* terms (I, me, my, we, us and our) than *out-group* terms (they, them and their). She also found posts to be written in 'short, paratactic sentences' employing 'informal, non-standard constructions and slang' (page 28). A further feature she identified as common to many weblogs was the use of frames (cf. Brown and Yule 1983), in which the author assumes that their audience shares their own background knowledge in the concepts they discuss. This supports the findings of Nowson et al. (2005) that personal weblogs have a degree of contextuality that many other genres of writing do not contain.

Personality and Blogging

Personality as described here refers to trait theory – where personality is decomposed into a number of measurable factors or traits. In this study we adopt the five-factor model of personality (cf. Costa and McCrae 1992). Table 1 describes behaviors associated with high and low scorers on each trait.

Table 1: Five traits of personality and their associated characteristics.

Trait	High behaviours	Low behaviours
Neuroticism	emotional instability; anxious; hostile; prone to depression	emotional stability; calm; less easily upset
Extraversion	extraverts; warmer; more assertive; action-oriented, thrill-seeking	introverts; low-key; deliberate; require less stimulation
Openness	appreciation for art and ideas; imaginative; more aware of feelings	more straightforward interests; conservative; resistant to change
Agreeableness	compassionate; cooperative; considerate; friendly	suspicious; unfriendly; wary; antagonistic; uncooperative
Conscientiousness	disciplined; dutiful; persistent; compulsive perfectionist	spontaneous; impulsive; less driven by desire for achievement

Given the psychological profiles proposed by Trammell et al. (2006, derived from Cywińska-Milonas, 2003), we may expect Extraverts to be more highly represented amongst the blogging population, with the auto-therapeutic aspect attracting more bloggers scoring higher on the Neuroticism scale. In fact, a previous study of blogger personality found a normal distribution for Neuroticism, Extraversion, Agreeableness and Conscientiousness (Oberlander and Nowson, 2006). Openness however showed a skewed distribution indicating that bloggers in their sample demonstrated higher levels of Openness.

Therefore rather than bloggers merely representing a particular subset of personality space of the population as a whole, we expect that it is the case that bloggers with different personalities write blogs serving differing purposes. For example, Extraverts may choose to document their lives using blogs, whereas for high Neurotic individuals, it is more likely to serve a cathartic or auto-therapeutic purpose. In the case of Openness and Conscientiousness, it is likely that their interests, opinions, or even feelings (but to a lesser extent than in the cathartic-style blogs) will be reflected in what they blog about: for the former arts and intellectual topics, whereas the latter may be more concerned with the daily grind. In the case of

Agreeableness, high scorers on this trait may be distinguished by what they *do not* write about, preferring instead to stick to 'safer' topics. Before presenting our hypotheses predicting how we expect these blog types to be characterized linguistically, we note previous findings relating personality to language.

Personality and Language

In monological writing situations (essay-writing and composing emails to friends), high Extraverts use more social words, references to themselves and others, positive emotion words, and express more certainty (Oberlander and Gill, 2006; Pennebaker and King, 1999). In email, they also show greater complexity, including increased use of introducing clause-initial connectives, such as *then*, *which*, and *what*, conjunctions and adjectives (Oberlander and Gill, 2006). Additionally, in blogs high Extraverts use more present tense verbs and talk of communication (Nowson, 2006). Conversely low Extraverts are likely to be characterized by more negations and negative emotion expressions, exclusive, inclusive, and causation words, and articles (Pennebaker and King, 1999), and by greater tentativeness (Oberlander and Gill, 2006), and in blogs talk more about achievements and use words relating to discrepancies (Nowson, 2006).

Turning now to Neuroticism, high scorers in monologue situations have been found to use more first person singular and negative emotion words (Pennebaker and King, 1999), and greater talk of discrepancies, jobs, and physical states (Nowson, 2006). Additionally, they use less outward-looking discourse containing fewer phrases referring to others, more exclusive and inclusive connectives, and a greater use of multiple-punctuation expressions (Oberlander and Gill, 2006). Low Neurotics on the other hand refer more to other people (Nowson, 2006), and use more nouns and adverbs (Oberlander and Gill, 2006).

In essays, high Openness scorers use more articles, longer words and insight words, and fewer first person singular, present tense, and causation words (Pennebaker and King, 1999); in blogs they again use more longer words, and also express positive feelings and use inclusive words. They also use fewer negations and write less frequently about school (Nowson, 2006).

High Agreeableness scorers use more first person singular and positive emotion words, and fewer articles and negative emotion words in essays (Pennebaker and King, 1999), and in blogs use fewer discrepancies and write less about body states (Nowson, 2006).

High Conscientiousness scorers use more positive emotion words, and fewer negations, negative emotion, causation, exclusive words, and discrepancies (Pennebaker and King, 1999), in blogs they are less likely to refer to topics concerned with death (Nowson, 2006).

Hypotheses

Previous investigation of the motivation and content of

blogs noted that self expression is the most likely motivation (reflected in content concerned with feelings or thoughts, rather than hobbies or interests; Trammell et al. 2006). However, here we investigate how such blog characteristics reflect personality, and also whether previous findings relating to personality and language content hold in this environment. We use an automated content analysis technique, focusing specifically on topic-related categories (LIWC, Pennebaker and Francis, 1999), rather than hand coding (Trammell et al. 2006).

Blogs are often characterised as apparently discrete categories (Nardi et al. 2004; Herring et al. 2005). However, given their relatively undefined and unstructured nature, we expect bloggers with different personalities to display more characteristics of one blog type than another. We note predictions by personality type.

1. Neuroticism – Blogs authored by high Neurotics are more likely to serve a cathartic or auto-therapeutic purpose. This is likely to overlap with findings from previous literature, namely, (a) greater self reference (first person singular pronoun) and negative emotion words (Pennebaker and King, 1999), and (b) fewer references to others (second, or third person pronouns) (Oberlander and Gill, 2006). Additionally from previous blog findings, we expect (c) the topic to focus more on jobs and physical states (Nowson, 2006),

2. Extraversion – We expect high Extravert blog authors to write blogs more concerned with documenting life, with this characterised by (a) more verbs (past, present and future), and time references. Consistent with previous literature, it is likely that Extraverts will use (b) more pronouns (first, second and third person) (Pennebaker and King, 1999). Additionally (c) we expect fewer negative emotion words (Pennebaker and King, 1999).

3. Openness – High Openness bloggers are likely to write blogs reflecting their interest, opinions or feelings. We therefore expect (a) topics to focus on leisure activities, and (b) a greater number of cognitive mechanism words and words concerned with the senses. From previous literature, we expect (c) fewer first person singular pronouns and present tense verbs (Pennebaker and King, 1999), and fewer references to occupation and more positive emotion words (Nowson, 2006).

4. Conscientiousness – We expect highly Conscientious bloggers to write about their interests and to (a) use more words relating to their occupation, and also to time, and past, present and future verbs. We also expect them to (b) use more positive emotion words and fewer negative emotion words (Pennebaker and King, 1999).

5. Agreeableness – We predict that this trait will mainly influence what topics the author chooses to write about or avoid in their blog. Following previous literature, we expect (a) fewer negative and more positive emotion words, and more self references (Pennebaker and King, 1999). We also expect (b) fewer bodily references (Nowson, 2006).

Method

Large scale blog data from authors with known personality was collected by first identifying an internet meme that was equivalent to a personality test. The items are simple yes/no responses across 5 questions per trait. Scores are calculated, though only the classes are shown: low (0-1 point); medium (2-3 points); and high (4-5 points). Despite the unknown origins and non-validated nature of the questionnaire, the items appear to be fairly standard markers for the big 5 model of personality (Buchanan, 2006; cf. Costa and McCrae, 1992). All bloggers who took this test were identified by their link to the source. Having identified URLs blog text for February to April was acquired from Nielsen BuzzMetrics' blog data. We note that part of this data set (February) was previously used in Nowson and Oberlander (2007).

Since this corpus contains thousands of bloggers and many more posts, we use an automated process for cleaning up the data. Processing the data is necessary in order to remove text not written by the author. The texts were cleaned up as follows:

- To remove memes which cause a significant portion of noise in personal blog text, anything within a post between <table> tags is removed (since the majority of memes are generated in this way).
- Similarly, any quoted text within <blockquote> tags was removed.
- All other HTML tags were then removed.
- Memes taking the form of lists (e.g., Top Ten...) were removed by identifying sequences of four or more consecutive numbers within a text.

We note that although this is likely not as accurate as human data cleaning, it is consistent, and the process can be replicated in future.

Additionally in processing the data, we control for the length of the texts in order that one individual, or group, do not exert undue influence. Therefore, while we only include bloggers who produced over 1000 words per month, more verbose individuals have their blogs capped at a 5000 word limit. This results in 14.8 million words, 5042 month blog collections (mean=2941.4 words, SD=1374.6), from 2393 participants, with each contributing an average of 2.11 months of blogs. Personality information for these participants is as follows (scale: low=1-high=3): Neuroticism=1.88 (SD=0.75); Extraversion=2.01 (SD=0.74); Openness=2.55 (SD=0.60); Agreeableness=2.22 (SD=0.69); Conscientiousness=2.23 (SD=0.67). As previously noted by Nowson and Oberlander (2007), we also find a positively skewed distribution for Openness.

When all month blog texts had been cleaned up, they were then processed using the LIWC content analysis tool (Pennebaker and Francis, 1999; as implemented in TAWC, Kramer, Oh, and Fussell, 2006). This is more appropriate than hand-coding given the size of our corpus (cf. Trammell et al. 2006). To capture the content of the blogs,

we use high level LIWC categories relating to our hypotheses, namely: cognitive mechanisms, sensory and perceptual processes, time, occupation, leisure activities, and physical states and functions. In addition, we also use the following finer grained categories: first person singular pronouns, second person pronouns, third person pronouns, positive emotion words, negative emotion words, and past, present, and future tense verbs. We note that second person pronouns are excluded from the current analysis since they did not meet the 1 percent usage criterion, which we adopt to ensure generalisability of our results (cf. Pennebaker and King, 1999). The social processes category was also omitted since it shows considerable overlap with the pronoun categories. Table 2 shows these categories and associated means and standard deviations.

Table 2: LIWC variables, means and standard deviations

LIWC variable	Examples	Mean (%)	Std. Dev.
1 st Person Pron.	I, my, me	6.66	2.34
2 nd Person Pron.	you, you'll	1.58	1.44
3 rd Person Pron.	she, their, them	2.03	1.19
Pos. Emotion	happy, pretty, good	3.01	0.91
Neg. Emotion	hate, worthless, enemy	1.79	0.71
Cognitive Proc.	cause, know, ought	6.22	1.38
Sensory Proc.	see, touch, listen	2.15	0.62
Time	hour, day, o'clock	4.54	1.18
Past Verbs	walked, were, had	4.22	1.59
Present Verbs	walk, is, be	10.34	2.35
Future Verbs	will, might, shall	1.19	0.52
Occupation	work, class, boss	1.91	0.73
Leisure Activ.	house, TV, music	1.20	0.60
Physical Conc.	ache, breast, sleep	1.75	0.76

Results

Results are reported for each personality trait in turn. In each regression model, we enter the personality trait as the dependent variable, with LIWC variables entered as independent variables. We note that since we treat the personality measures as ordinal data, we build ordinal logistic fit models for each trait. In each model 'month' of the blog is entered, to ensure that different month contributions by the same author are accounted for (not reported in the results). The resulting models are shown in Tables 4-8. Consulting these tables reveals that overall these linguistic features relate to a relatively low amount of variance in personality score, with Neuroticism showing the greatest (2.35%), and Conscientiousness the least (1.06%).

Table 4: Language features relating to blogger Neuroticism

Variable	X2	Sig (p)		
1 st Person Pron.	30.20	<.0001		
3 rd Person Pron.	- 22.01	<.0001		
Pos. Emotion	- 26.57	<.0001		
Neg. Emotion	114.49	<.0001		
Time	- 9.41	0.0022		
Occupation	- 4.86	0.0275		
Model R2	N	DF	X2	Sig (p)
0.0133	5042	10	254.02	<.0001

Neuroticism (Table 4) – consistent with our predictions of blogging as catharsis, and also with previous literature (H1a), we find that High Neurotics use more first person singular pronouns, and negative emotion words. H1b and H1c are partially supported, with High Neurotics as expected using fewer third person pronouns and fewer references to occupation, but showing no difference in use of second-person pronouns or physical references. In addition we also find that high Neurotics use fewer positive emotion words and fewer references to time.

Table 5: Language features relating to blogger Extraversion

Variable	X2	Sig (p)		
1 st Person Pron.	26.62	<.0001		
2 nd Person Pron.	23.96	<.0001		
3 rd person Pron.	14.29	0.0002		
Pos. Emotion	29.67	<.0001		
Neg. Emotion	8.06	0.0045		
Cognitive Proc.	- 23.25	<.0001		
Past Verbs	- 8.24	0.0041		
Occupation	- 10.00	0.0016		
Model R2	N	DF	X2	Sig (p)
0.0133	5042	9	142.71	<.0001

Extraversion (Table 5) – For high Extraverts, we unexpectedly (H2a) find a decrease in the use of past verbs, and find no difference in the use of other verbs or time references. However, consistent with previous literature we do find greater use of first, second, and third person pronouns (H2b). Contrary to our predictions (H2c) we actually found that high Extraverts actually used a greater number of negative emotion words. In addition to these predictions, we also found an increase in the use of positive emotion words, as well as a decrease in the use of cognitive mechanism and occupation words.

Table 6: Language features relating to blogger Openness

Variable	X2	Sig (p)		
1 st Person Pron.	6.87	0.0088		
Neg. Emotion	15.33	<.0001		
Time	- 26.33	<.0001		
Past Verbs	- 27.40	<.0001		
Present Verbs	- 10.06	0.0015		
Leisure Activ.	7.46	0.0063		
Physical Conc.	13.25	0.0003		
Model R2	N	DF	X2	Sig (p)
0.0163	5042	9	135.52	<.0001

Openness (Table 6) – High Openness bloggers as expected (H3a) wrote more about leisure activities, but showed no difference in their use of cognitive mechanism or sense words (H3b). For H3c as expected we find fewer present tense verbs, but contrary to expectation we also find an increase in use of first person singular pronouns, and no difference in the use of occupation or positive emotion words. In addition, we note an increase in high Openness scorer use of negative emotion words and physical concern words, and a decrease in the use of past tense verbs and time words.

Table 7: Language features relating to blogger Conscientiousness

Variable	X2	Sig (p)		
3 rd person Pron.	7.69	0.0056		
Pos. Emotion	4.53	0.0332		
Neg. Emotion	- 14.96	0.0001		
Time	7.31	0.0069		
Occupation	34.18	<.0001		
Physical Conc.	- 8.48	0.0036		
Model R2	N	DF	X2	Sig (p)
0.0106	5042	8	105.6397	<.0001

Conscientiousness (Table 7) – As expected (H4a), high Conscientiousness scorers use more words relating to their occupation and to time. However this hypothesis is only partially supported, since no differences in use of past, present or future verbs are found for this trait. For H4b, as predicted by the literature, we find that high Conscientiousness relates to greater use of positive emotion words and use of fewer negative emotion words. Additionally we find an increase in use of third person pronouns, and decrease in use of physical concern words.

Table 8: Language features relating to blogger Agreeableness

Variable	X2	Sig (p)		
1 st Person Pron.	36.75	<.0001		
Pos. Emotion	56.68	<.0001		
Neg. Emotion	- 90.49	<.0001		
Leisure Activ.	- 5.29	0.0214		
Model R2	N	DF	X2	Sig (p)
0.0155	5042	6	157.8489	<.0001

Agreeableness (Table 8) – As expected from previous literature (H5a), highly Agreeable authors used fewer negative emotion words and more positive emotion words; We found no difference in use of physical concern words (H5b), but additionally found an increase first person singular pronoun usage and also a reduction in talk about leisure activities.

Discussion

In this section we will review different characteristics of language use by personality groups, and note what these indicate about the motivation and topic of the bloggers (and relate this back to our original predictions). Additionally, we will also note instances where the current findings differ from those previously found for personality and language, since these may provide insight into how different personality groups modify their behavior in the blogging environment in order to present a particular image (Goffman, 1959).

As previously noted, the linguistic variance of topic in blogs relating to personality is small (around 1-2%) compared to other studies of blogs which include both stylistic and topic-related language features (11-41%; Nowson, 2006). This apparently indicates the value of including a combination of features relating to style and topic for author personality classification tasks, at least in blogs.

In terms of language use as an indicator of blogger motivation or blog type, we consider each personality trait in turn. In the case of Neuroticism, our predictions that they use blogging as catharsis or self-therapy are supported by the high Neurotic use of negative emotion words and references to themselves, and fewer references to others using third person pronouns. The unpredicted findings regarding the high Neurotic decreased use of positive emotion words and time references also apparently indicate a focus on venting purely negative feelings, and possibly a focus purely on the present or lack of putting events in a temporal context.

For Extraversion, we find that in general – as in other media – high Extraverts talk more about themselves and other people, and also address the reader directly. Contrary to findings in other contexts, Extraverts actually used a greater number of negative emotion words; however this occurs alongside an increase in positive emotion words. That high Extraverts use fewer cognitive- and job-related words, and demonstrate no difference in verb or time usage, contributes to the impression that they are not using the blog to document the mundane aspects of their existence. Taken together, these are consistent with Extraverts using blogging as a means to document their life, although this seems to revolve around interactions with the reader and other people, as well as themselves. The increase in positive and negative emotion words, seems to indicate that in this ongoing communication with the reader high Extravert bloggers share their emotional experiences (good and bad), perhaps in a way that they

would not normally do in other forms of communication. It should be noted that this is apparently different to the emotional venting of the high Neurotic bloggers who focus on writing about negative emotions, especially in the context of themselves and may in fact use blogging for auto-therapeutic purposes.

For Openness, we find that it is the topic of leisure activities (which further examination indicated was particularly content categories relating to the arts, such as music and TV) which most distinguishes bloggers high in Openness. It is however, surprising that words relating directly to thinking and intellectual activities (cognitive mechanisms), or senses showed no relationship to Openness, although we do note fewer present tense verbs, along with past tense words and time references perhaps indicating less concern with the here and now. Contrary to expectation, we found that high Openness scorers showed increased self-reference (first person singular pronouns), along with increased negative emotion and physical concern words. These latter characteristics appear similar to those found for high Neuroticism, however we note that although there is an increased use of self reference, there is not a corresponding reduction in references to others. Additionally, in the case of negative emotion words, further examination of the individual constituent categories (anxiety, anger, and sadness words), indicates that high Neurotics show increased use of all of these categories, whereas high Openness bloggers solely use more anger words. Overall these findings do to some extent support the proposal that high Openness scorers are more likely to write about leisure interests, rather than mundane topics, with their use of negative emotion words and first person pronouns perhaps indicating a review or evaluation component, not normally found in other media.

High levels of Conscientiousness in bloggers as predicted relates to more mundane activities, such as work and time (despite no differences in verb use), with previous findings from the literature supported in their use of positive, rather than negative emotion, words. Given their increased use of references to others, this appears to be more of a commentary on life going on around them, which can be contrasted with the documentation of their own life and venting of the high Extraverts, the high Neurotic catharsis, and the commentary and review of leisure topics by the high Openness scorers.

In terms of Agreeableness, as predicted authors scoring high on this trait use fewer negative, and more positive, emotion words. Additionally, in the context of blogs our findings suggest that physical concerns are a topic that should not necessarily be avoided, with self-reference viewed positively, but leisure talk viewed less so.

To summarize, in this study, we have analyzed content features of a large blog corpus to investigate the role of personality in blogger motivations. As predicted, our findings support the hypothesis that highly Neurotic authors use blogging for cathartic or auto-therapeutic purposes. Whilst confirming that high Extraverts use blogs to document their life, this tends not to focus on the

mundane minutiae, but as in other situations, they talk more about themselves and other people. Interestingly, we find that they engage the reader directly using second person pronouns (e.g., “you”), a behavior apparently distinctive of this trait. Unexpectedly, high Extraversion is linked to the expression of positive and negative emotion, indicating an additional venting purpose to the blogs.

In the case of high Openness in bloggers, this was predicted to be reflected in interests, opinions and feelings. Leisure interests (which we suppose to be music and TV) did indeed relate to Openness; however we expected greater emphasis upon the intellectual or sensing aspect of this trait. Instead, the blogs of high Openness scorers appear to represent more of an evaluation or review with greater self-reference and negative emotion words (sadness). Like Openness, Conscientiousness influenced blogs in terms of topics of interest to the author; however in the case of Conscientiousness this appears to be more of a mundane commentary regarding life around them. For Agreeableness, we note that as elsewhere, positive rather than negative emotions are more pleasing to others, but that in blogs, self talk is apparently more acceptable, whereas talk about leisure activities is less so. In all of these cases, however, we must emphasise the relatively small amount of variance explained by personality. Indeed, it may be the case that some traits are better expressed through an individual’s physical (or potentially virtual) environment (Openness), or through what they spend their time doing and where they do this (Conscientiousness, Agreeableness) (cf. Gosling, Gaddis and Vazire, 2008; Mehl, Gosling, and Pennebaker, 2006).

Finally, we note the role of blogging as a form of expression. Although blogs afford their author freedom of topic and style, in most cases the bloggers’ behavior was consistent with our predictions based on their personality type or their linguistic behavior in other media. This apparently suggests that bloggers *do not* generally choose to present themselves differently to the way they present in other contexts (cf. Goffman, 1959). However, we do note that the blog media does apparently lead to a certain amount of adaptation, for example with high Extraverts using it to express emotions, and high Openness authors providing reviews or evaluations of topics of interest, or of the greater prevalence of self-reference observed among highly Agreeable authors. We propose future work to explore personality and self-reported blogger motivation, as well as author reports of self-presentation online.

Conclusion

In this paper, we have used a large blog corpus to examine the role of personality in motivation for blogging. Linguistic content of the blogs revealed cathartic and auto-therapeutic tendencies of high Neurotic bloggers; high-level life documentation and emotion expression by high Extraverts; commentary and evaluation by high Openness scorers, and reports of daily life by high Conscientiousness bloggers. Highly Agreeable authors demonstrate similar

strategies to those they adopt in other contexts. In general, these findings are consistent with other contexts, and indicate that bloggers tend to adapt to the possibilities of the medium, rather than try to present themselves differently.

Acknowledgments

We are grateful to Darren Gergle and Lauren Scissors (Northwestern University) for comments on earlier versions of this work, and to Iain MacMillan (Ryan MacMillan Ltd., London) for informative discussions relating to blogging and online influence.

References

- Brown, G. and Yule, G. 1983. *Discourse Analysis*. Cambridge: Cambridge University Press
- Buchanan, T. 2006. Personal Correspondence, received 23/11/06.
- Cohn, M., Mehl, M.R. and Pennebaker, J.W. 2004. Linguistic markers of psychological change surrounding September 11, 2001. *Psychological Science*, 15, 687-693.
- Costa, P., and McCrae, R.R.. 1992. *Professional Manual. Psychological Assessment Resources*, Odessa, FL.
- Efimova, L., and de Moor, A. 2005. Beyond personal webpublishing: An exploratory study of conversational blogging practises. *Proceedings of the 37th Annual HICSS Conference*. Big Island, Hawaii.
- Glance, N., Hurst, M., and Tomokiyo, T. 2004 BlogPulse: Automated Trend Discovery for Weblogs. in *Proceedings of WWW 2004*. New York, US.
- Goffman, E. 1959. *The Presentation of Self in Everyday Life*. Double Day: Garden City, New York.
- Gosling, S. D., Gaddis, S., and Vazire, S. (2008). First impressions from the environments that we create and inhabit. In J. Skowronski, J., and Ambady, N. (Eds.), *First Impressions*. (pp. 334-356). New York: Guilford.
- Gruber, H. 2000. Scholarly Email Discussion List Postings: a single new genre of academic communication? In Pemberton, L. and Shurville, S. (Eds) *Words on the Web*, 36-43. Exeter: Intellect Books.
- Herman, D., Jahn, M., and Ryan, M.-L. (eds.). 2005. *The Routledge Encyclopedia of Narrative Theory*. London, Routledge.
- Herring, S., Scheidt, L., Sabrina Bonus, S., and Wright, E.

2005. Weblogs as a bridging genre. *Information, Technology & People*, 18 (2), 142-171.
- Heylighen, F., and Dewaele, J.-M. 2000. Variation in the contextuality of language: an empirical measure. *Foundations of Science*, 7, 293-340.
- Kelleher, T., and Miller, B.M. 2006. Organizational blogs and the human voice: Relational strategies and relational outcomes. *Journal of Computer-Mediated Communication*, 11(2).
- Kramer, A.D.I., Oh, L.M., and Fussell, S.R. 2006. Using linguistic features to measure presence in computer-mediated communication. *CHI 2006 Notes* (pp. 913-916). NY: ACM Press.
- Li, D. 2005. Why do you blog: A uses-and-gratifications inquiry into bloggers' motivations. Unpublished Masters Thesis, Marquette University.
- McNeill, L. 2005. Genre Under Construction: The diary on the internet. *Language@Internet*. <http://www.languageatinternet.de/articles/2005/120/>
- Mehl, M. R., Gosling, S. D., and Pennebaker, J. W. (2006). Personality in its natural habitat: Manifestations and implicit folk theories of personality in daily life. *Journal of Personality and Social Psychology*, 90, 862-877.
- Mishne, G. 2005. Experiments with Mood Classification in Blog Posts. In *Style2005 - the 1st Workshop on Stylistic Analysis Of Text For Information Access*, at SIGIR 2005.
- Nallapati, R., and Cohen, W. 2008. Link-PLSA-LDA: A new unsupervised model for topics and influence of blogs. *International Conference on Weblogs and Social Media 2008*. Seattle, WA.
- Nardi, B.A., Schiano, D.J., Gumbrecht, M., and Swartz, L. 2004. Why We Blog. *Communications of the Association for Computing Machinery*. December. Pp. 41-46.
- Nilsson, S. 2003. The function of language to facilitate and maintain social networks in research weblogs. Dissertation, Umea Universitet, Engelska Lingvistik.
- Nowson, S. 2006. The Language of weblogs: a study of genre and individual differences. Unpublished Doctoral Thesis, The University of Edinburgh
- Nowson, S., and Oberlander, J. 2007. Identifying more bloggers: Towards large scale personality classification of personal weblogs. *International Conference on Weblogs and Social Media 2007*. Boulder, Colorado.
- Nowson, S., Oberlander, J., and Gill, A.J. 2005. Weblogs, Genres and Individual Differences. In *the proceedings of the 27th Annual Conference of the Cognitive Science Society*. Stresa, Italy.
- Nowson, S. forthcoming. The language of weblogs and weblog authors. In Herring, S.C., Stein, D., and Virtanen, T. (Eds), *Handbook of the Pragmatics of Computer-Mediated Communication*.
- Oberlander, J., and Gill, A.J. 2006. Language with character: A corpus-based study of individual differences in e-mail communication. *Discourse Processes*, 42(3), 239-270.
- Oberlander, J., and Nowson, S. 2006. Whose thumb is it anyway? Classifying author personality from weblog text. In *the proceedings of the Association for Computational Linguists 2006*. Sydney, Australia.
- Pennebaker, J.W., and Francis, M.E. 1999. *Linguistic Inquiry and Word Count: LIWC*. Mahwah, NJ: Erlbaum.
- Pennebaker, J.W., and King, L. 1999. Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology*, 77, 1296-1312.
- Puschmann, C. Forthcoming. "Thank you for thinking we could" - Use and function of interpersonal pronouns in corporate web logs. Dorgeloh, H. and Wanner, A. (eds.): *Approaches to Syntactic Variation and Genre*. Berlin/New York: Mouton de Gruyter.
- Sorapure, M. 2003. Screening moments, scrolling lives: Diary writing on the web. *Biography: An Interdisciplinary Quarterly*, 26 (1), 1-23.
- Trammell, K.D., Tarkowski, A., Hofmokl, J., and Sapp, A.M. 2006. Rzeczpospolita blogów [Republic of Blog]: Examining Polish bloggers through content analysis. *Journal of Computer-Mediated Communication*, 11(3), article 2. <http://jcmc.indiana.edu/vol11/issue3/trammell.html>
- Yates, S.J. 1996. Oral and Written Linguistic Aspects of Computer Conferencing: A Corpus Based Study. In Herring, S.C. (Ed.) *Computer Mediated Communication: Linguistic, social and cross-cultural perspectives*, 29-46. Amsterdam: John Benjamin.