Weblogs, Genres, and Individual Differences

Scott Nowson (S.Nowson@ed.ac.uk)
School of Informatics; University of Edinburgh
2 Buccleuch Place, Edinburgh, EH8 9LW, UK

Jon Oberlander (J.Oberlander@ed.ac.uk)
School of Informatics; University of Edinburgh
2 Buccleuch Place, Edinburgh, EH8 9LW, UK

Alastair J. Gill (A.Gill@ed.ac.uk)
School of Informatics; University of Edinburgh
2 Buccleuch Place, Edinburgh, EH8 9LW, UK

Abstract

Blogs are personal online diaries, and a relatively recent form of computer-mediated communication. What kind of writing do they contain? This paper adopts a measure of linguistic contextuality/formality, due to Heylighen and Dewaele, and applies it to a corpus of weblogs. It first compares the corpus with sub-corpora from the British National Corpus, and weblogs are shown to be more formal than e-mail, but less formal than biographies. Then, the paper explores the impact of individual differences between writers on their texts' contextuality/formality. It appears that Extraversion and Neuroticism are less influential than previously supposed, and it is argued that gender and Agreeableness account for more of the variability in the extent to which weblog writers take their readers' contexts into account.

Blog It All

27% of internet users read weblogs, but 62% still don't know what they are. This is in spite of the fact that an estimated 8 million weblogs have been created in the US alone (Rainie, 2005). A weblog is a frequently updated website which contains news and views on a variety of topics, from politics to gossip. Weblogs are already being seen as a powerful news-gathering medium (Belo, 2004). The term 'blog' is more widely used, and is most commonly used to refer to the sub-category of online personal diaries. Mirriam-Webster named 'blog' as their Word of the Year 2004.

The internet is increasingly considered as a resource for linguistic study (Keller, Lapata and Ourioupina, 2002; Volk, 2001), and a number of studies have focused on the nature of various types of computer-mediated communication (CMC), such as asynchronous e-mail and synchronous chat. A key reason for studying these genres is to determine how they differ from non-computer-mediated analogues. In addition, however, CMC has the great virtue that it can make available large amounts of naturalistic language use, at relatively low cost; in particular, transcription costs are small compared to those associated with spontaneous speech. As a result, CMC environments offers an excellent arena in which to study the projection and perception of individual differences (Hancock and Dunham, 2001; Markey and Wells, 2002; Gill and Oberlander, 2002, 2003). To date, however, relatively little work has considered blogs though they are now being discussed as both a topic (Cohn, Mehl, & Pennebaker, 2004, Rosenbloom, 2004) and a tool for study (Mortensen and Walker, 2002). This paper discusses the place of blogs within a range of genres, and investigates some aspects of individual differences between writers.

A corpus was gathered by asking authors of personal weblogs—'bloggers'—to complete a sociobiographic questionnaire and an on-line implementation of an IPIP Five Factor Personality Inventory (41 items scored from a 5 point scale, Buchanan, 2001). A corpus of text was created by asking each blogger to submit text they had previously written, for a whole month preceding the date of the questionnaire. The month was specified so as to reduce the effects of subjects choosing their 'best' or 'favourite' month. The resulting corpus consisted of 71 subjects (24 males and 47 females), with ages ranging from 15-50 (mean 28.4). When all text considered 'personal,' that which discussed individual concerns rather than general ones, was collected from the larger HTML files, it amounted to over 410,000 words.

The rest of this paper is structured as follows. The next section discusses previous work on individual differences in language production, following Dewaele and Furnham. It then introduces a unitary measure of a text's formality (as opposed to contextuality): the F-measure, due to Heylighen and Dewaele. The section following this considers blogs as a genre, using the F-measure to compare the corpus with sub-corpora from the British National Corpus. The paper then turns to the impact of individual differences between writers on their texts' contextuality/formality. The paper concludes by considering which dimensions of individual variation are most worthy of further study.

Background

Extraversion and Neuroticism are the two dimensions of variation which are common to the two major trait theories of personality: Costa and McCrae's five factor model (Costa and McCrae, 1992); and the three factor model of Eysenck (Eysenck and Eysenck, 1991). Level of Extraversion is associated with how outgoing and assertive a person is; level of Neuroticism is associated with how anxious, self-conscious and temperamental a person tends to be (Matthews, Deary and Whiteman,

2003). Perhaps because of its connection to communicativeness, the Extraversion dimension has been studied particularly intensively in work on the relationship between personality and language use.

Furnham (1990) proposed the following description of Extravert language: it is less formal; has a less restricted code; it uses vocabulary more loosely, and uses more verbs, adverbs and pronouns. Using factor analysis of syntactic tokens produced by L2 speakers, Dewaele and Furnham (2000) describe *implicit* language as a preference for pronouns, adverbs and verbs, while *explicit* language involves a preference for nouns, modifiers and prepositions. They find that Extraverts prefer implicitness, while Introverts prefer explicitness.

Oberlander and Gill (2004) investigated this 'implicit-extravert hypothesis' on a corpus of e-mail texts collected from 105 subjects, who had been asked to write two e-mails *to a good friend whom they hadn't seen for quite some time*. However, Pennebaker and King (1999)'s factor analysis of a collection of texts from writers of known personality had previously isolated a factor for 'Immediacy', which correlated with high Neuroticism. Gill and Oberlander (2003) also found that high Neurotics had a preference for common words that occurred frequently in speech. This led Oberlander and Gill to investigate an 'implicit-neurotic hypothesis', whereby high Neurotics would prefer implicit expressions, and low Neurotics explicit. They found that only some parts of speech were used with significantly different relative frequencies by the sub-groups of their subjects. However, n-gram analysis for part-of-speech sequences revealed some support for both hypotheses. One question raised by that study is whether a larger corpus would provide clearer evidence of implicitness effects. Another question is whether there is a better way of measuring a text's implicitness.

The second question is addressed by Heylighen and Dewaele (2002). They explore the notion of implicitness in greater detail and develop a unitary measure of a text's relative contextuality (implicitness), as opposed to its formality (explicitness). Briefly, they consider a notion of *deixis* following Levelt (1989), and demarcate a group of expressions that must be anchored to some part of the spatio-temporal context of utterance in order to be properly interpreted. Greater use of these expressions leads to greater *contextuality*, while greater use of non-deictic expressions leads to greater *formality*. They propose that certain parts of speech (such as verbs) are generally (although not invariably) deictic in nature, while others (such as nouns) are generally non-deictic. They then define the F-measure, a single measure of a text's contextuality versus formality: a low score indicates contextuality, represented by a greater relative use of pronouns, verbs, adverb, and interjections; a higher score indicates formality, represented by greater relative use of nouns, adjectives, prepositions and articles. F is defined as follows:

$$F = 0.5 * [(nounfrq + adjfrq + prepfrq + artfrq) - (pronfrq + verbfrq + advfrq + intfrq) + 100)]$$

Heylighen and Dewaele tested the idea via factor analysis of part-of-speech data and found that over 50% of the variance was accounted for by a factor very similar to their measure. They used the F-measure to explore corpus data derived from Dutch, Italian, and English sources. The results were consistent: spoken language scored lower than written language, meaning that the latter is more formal; newspapers were more formal than works of fiction; interview data was more formal than casual conversation.

Importantly, Heylighen and Dewaele were also able to analyse the relation between gender and contextuality, and the results consistently showed that women used more contextual language—both in written and spoken texts—while men tended toward formality.

Of course, there are other factors which can be used to distinguish between genres. Following an extensive factor analysis of 67 linguistic features, Biber (1988) found a number of significant factors. One of these factors, termed 'involved versus informational production' concerned amongst others, most of the variables in the F-measure. Loewerse, McCarthy, McNamara and Graesser (2004) followed Biber by repeating his analysis with a new set of 236 language and cohesion features, including a number of LSA-based metrics. They too found several factors that readily highlight differences in genres.

However, it is Heylighen and Dewaele's F-measure which has been used specifically to investigate individual differences between writers *within* a genre, so we adopt their measure here. Along side gender effects, following Dewaele and Furnham's earlier work on personality and implicitness, it is expected that there should be a correlation between Extraversion and contextuality: Extraverts will prefer the use of the contextual parts of speech, and Introverts will prefer their more formal counterparts.

Differences Between Genres

In order to understand blogs as a genre, it is necessary to place them in a larger context. To this end, we chose to compare the genre to a range selected from the British National Corpus (BNC). The BNC consists of over 4000 files, containing over 100 million words of both spoken and written English. Calculating the F-score of a number of genres from the BNC allows us to place blogs on a scale and furnishes an opportunity to test the face validity of Heylighen and Dewaele's F-measure by examining the plausibility of that scale. The F-score of Gill and Oberlander's e-mail corpus can also be calculated, and included in the placement.

Method

Using Lee's BNC World Edition Index[1] (2001), 17 genres were selected from the BNC. These included both spoken ($n = 4$) and written ($n = 13$) material, ranging from sermons and fiction writing, to text taken from newspapers and academic works. Only files dating from 1985 to 1994 and (for speech) only spoken files with a single speaker were included. Altogether there were 837 files comprised of 23 million words. The original release

[1]Available at http://clix.to/davidlee00

Table 1: Average F-score of selected genres from BNC

Genre	Ave F
Sermons	42.4
Lectures on Social Science	44.3
Unscripted Speeches	44.4
Fiction Prose	46.3
Personal Letters	49.7
Sports Mailing List E-Mails	50.0
Scripted Speeches	53.0
School Essay	53.2
Biography	56.3
Non Academic Social Science	56.9
Nat Broadsheet Social	57.5
Professional Letters	57.5
Nat Broadsheet Editorial	58.1
Nat Broadsheet Science	60.0
University Essays	60.3
Academic Social Science	60.6
Nat Broadsheet Reportage	62.2

of the BNC comes pre-tagged using the CLAWS tagset. These tags are algorithmically reduced to the set needed for calculating the F-score of each file. These scores are then averaged to give the F-score of each genre.

Both the blog and e-mail corpora have also been tagged using the MXPOST tagger (Ratnaparkhi, 1996) and the PENN tagset. These tags were mapped down to the same set for comparison. Each e-mail file contained 2 messages from the same writer ($n = 105$) while each blog file contained all the text for an author from one month ($n = 71$).

Results

When the F-score calculation was completed on the BNC genres selected, they ranked as in Table 1. As predicted by Heylighen and Dewaele (2002), spoken genres are on the whole less formal than written, with sermons, lectures, and unscripted speeches scoring the lowest. Scripted Speeches are more formal than Unscripted and also those written genres considered least formal: Fiction, Personal Letters and E-Mails. Many of the results are intuitive: Academic writing is more formal than Non-Academic; Professional Letters are more formal than Personal; University-level Essays are more formal than School level. We also see degrees of similarity: Personal Letters are close to the BNC's E-Mails (which come from a mailing list; cf. Collot and Belmore, 1996).

The F-score was calculated for the new blog corpus and Gill and Oberlander's existing e-mail corpus. The results are displayed, along with those of the closest genres selected from the BNC, in Table 2. As one might expect, the e-mail corpus is very similar to the E-Mails taken from the BNC; proximity to Personal Letters follows from this. It can be seen that the blogs are scored as being significantly less contextual than the e-mails (t=3.54, DF=174, p<.001).

Table 2: Average F-score of E-Mail and Blog corpora as situated in the BNC genre ranking.

Genre	Ave F
Sports Mailing List E-Mails	50.0
E-Mail Corpus	*50.8*
Scripted speeches	53.0
School Essay	53.2
Blog Corpus	*53.3*
Biography	56.3

Discussion

This particular result can be explained by considering some of the situational factors involved in deixis. Heylighen and Dewaele draw on four categories: the *persons* involved, the *space* of the communication, the *time*, and the prior *discourse*. When collecting e-mail data, subjects were instructed to imagine they were writing to a friend—a single person who knew them. The blog data however, was collected from web-published blogs. These can be read by persons unknown to the writer; hence, to some extent, they are written with such readers in mind. So bloggers cannot assume as large a shared context with their readers as writers of e-mails composed for friends.

Not knowing the reader means the writer can assume less about any knowledge of any places, or *spaces* that are discussed. Similarly, since one cannot know when a blog post will be read, or whether any previous posts have been read, the writer can assume less about the *time* and *discourse* contexts.

In sum, it appears that the F-measure of contextuality is a reasonable method for distinguishing *between* genres. In fact, the ordering on genres is very similar to that found by Biber (1988) when ranking via his involved/informational factor. However, as noted above, the current measure of contextuality/formality can also be used to explore individual differences between writers *within* a genre.

Individual Differences Within Genres

The individual differences under investigation here mainly concern those of personality. The hypotheses are that the F-measure correlates negatively with both Extraversion and Neuroticism. But, following Heylighen and Dewaele, and to further test the validity of their measure, we can first test for gender differences.

Gender differences

Gender has previously been investigated in the BNC, for instance in the Conversational sub-corpus looking at a word level (Rayson, Leech and Hodges, 1997), and in written work using sub-word level characteristics (Argamon, Koppel, Fine and Shimoni, 2003).

Heylighen and Dewaele applied their F-measure to texts of known gender and found a distinct difference between the sexes. Females score lower, preferring a more contextual style, while men prefer a more formal style.

Table 3: Average F-score for Male and Female writer in selected genres

Genre	Male	Female
Fiction prose Adult	47.8	45.0
E-Mail Corpus	*53.1*	*49.5*
Blog Corpus	*55.2*	*52.4*
Non academic Social Science	59.5	52.1
Academic Social Science	60.5	60.8

Table 4: Pearson Correlation between F-score and personality trait.

Trait	r
Neuroticism	−.090
Extraversion	−.098
Openness	.162
Agreeableness	.272*
Conscientiousness	.028

Note: two-tailed, $^*p<0.05$

This result was taken to be consistent with previous findings from socio-linguistic and psychological studies.

A number of the genres selected from the BNC are marked for author gender, as are the e-mail and blog corpora. Table 3 shows the average F-score for males and females in the genres for which data was available. For both genders, the ordering of genres remains as shown in Tables 1 and 2. Females score lower in four out of five genres. Within the blog corpus this difference is significant (t=2.90, DF=69, p<.005). The exception is when the writing is academic in nature. Here there is little difference between male and female F-scores; both are relatively high. It appears that while females prefer a more contextual style, when required, they can adopt a style at least as formal as that projected by males.

Personality correlation

A starting point for personality analysis is to test the correlation between F-score and writer trait score. While Extraversion and Neuroticism have already been discussed, the remaining traits of the Five Factor Model are Openness, Agreeableness and Conscientiousness. Openness is characterised by culture, intellect and originality; level of Agreeableness is associated with how compliant, straightforward and altruistic a person is; Conscientiousness concerns how competent, deliberate and self-disciplined an individual is (Matthews *et al.*, 2003). The results of the Pearson Correlation analysis for all files in the blog corpora along the Five Factor dimensions are displayed in Table 4. The results are not as expected. Given the implicit-extravert and implicit-neurotic hypotheses, we expect a negative correlation with Neuroticism and Extraversion. The correlation is in the expected direction, but it is small, and does not reach significance. However, we do find a stronger, positive, and significant correlation with Agreeableness. The correlation with Openness is also reasonably strong and positive, but does not reach significance. Conscientiousness shows the smallest correlation of all.

Table 6: Average F-score of corpus stratified by trait

Trait	Low	Mid	High
Neuroticism	54.1	53.0	53.8
Extraversion	54.8	53.0	53.2
Openness	52.7	53.1	54.4
Agreeableness	52.3	53.1	55.9

To gain a better perspective on what is happening, we can also look at the frequencies of the individual parts-of-speech that define the F-measure. When there is an overall negative correlation between trait score and the F-score—as with Extraversion and Neuroticism—we might expect a negative correlation between trait score and frequencies for nouns, adjectives, prepositions and articles, while there should be a positive correlation for pronouns, verbs, adverbs and interjections. The opposite should hold when there is positive correlation between trait score and the F-score—as with Agreeableness and Openness. Table 5 displays the results. As might be expected from the overall correlations shown in Table 4, we here find that Agreeableness has the strongest correlations, and the most that reach significance. Openness also has some reasonably strong and significant correlations. None of the Extraversion and Neuroticism correlations reach significance. And once again, there are only very small correlations for Conscientiousness.

However, with only a couple of small exceptions, the directions of the correlations are as expected. Neuroticism and Extraversion scores correlate positively with the frequencies of contextual parts of speech, and negatively with those parts of speech considered formal. The opposite is true for Agreeableness and Openness.

Stratified corpus analysis

It therefore appears that there is some relation between contextuality/formality, for the four personality dimensions of Neuroticism, Extraversion, Openness and Agreeableness. But the relation is stronger in some cases than in others. To take a closer look at each case, we adapted the stratification approach used in Oberlander and Gill (2004). Here, High and Low personality sub-groups are created for each personality dimension by splitting off the groups at greater than 1 standard deviation above, and below, the mean score for each dimension. The remainder of the subjects are allocated into the Mid sub-group for that dimension. For the current analysis, we have therefore dropped the further requirement that writers in a given sub-group for a dimension had to have scores *within* 1 standard deviation of the mean on the other dimensions. With this simpler stratification strategy, we retain all 71 subjects for the exploratory analysis. The average F-score for the sub-groups, by dimension, can be seen in Table 6. As might be expected

Table 5: Pearson Correlation between POS frequency and personality trait.

Trait	Noun	Adjective	Prep'n	Article	Pronoun	Verb	Adverb	Interj'n
Neuroticism	–.117	.128	–.075	–.077	–.013	.150	.127	.016
Extraversion	.017	–.092	–.117	–.148	.231	.024	–.044	.110
Openness	.055	.354**	.268*	–.002	–.005	–.152	–.244*	–.221
Agreeableness	.196	.165	.151	.260*	–.173	–.257*	–.263*	–.240*
Conscientiousness	.053	.007	–.054	.041	–.095	–.053	–.010	–.023

Note: two-tailed, *p<0.05, **p<0.01

Table 7: Multiple regression of the F-score

Dependent variable	Independent variable	β	R^2	p
F-score	Gender	.33		
	Agreeableness	.27	**.18**	**.001**

from the overall correlations, the F-scores for Low Neurotics and Low Extraverts are greater than the F-scores for High sub-groups. The opposite trend holds for the other dimensions. Moreover, on Openness and Agreeableness, the F-scores for the Mid sub-groups are intermediate between the F-scores for the Low and High sub-groups. However, for Neuroticism the F-scores for the Mid sub-group are lower than those of either end sub-group. There is thus a hint of non-linearity in these results. This is consistent with Gill, Harrison and Oberlander's (2004) finding that, for inter-personal priming, High and Low sub-group performances resemble one another more than they resemble that of the Mid sub-group. The result for Extraversion is harder to accomodate; the closeness of the Mid and High sub-groups suggesting that only high Introversion affects formality.

Regression analysis

A final confirmatory analysis involved determining which of the personality variables, if any, best accounted for variance in F-score. To this end, a step-wise multiple regression was employed. The F-score was considered the dependent variable; the personality traits, along with gender and age, the independent variables. The results are displayed in Table 7. Only 18% of the variance of the F-score is explained. Gender makes the most significant contribution. The only personality trait that enters the regression equation is Agreeableness, and this was indeed the personality trait that showed the highest correlation (see Table 4). Note that gender and Agreeableness are independent, Pearson's r=-.004, *ns*.

Discussion

Neuroticism and Extraversion did not correlate with contextuality/formality as strongly as we had expected. There was a small and non-significant effect in the expected direction. So, support for implicit-extravert and implicit-neurotic hypotheses is not forthcoming, although the results do seem consistent with the idea that, for this measure, the two traits lead to similar results. In fact, Dewaele and Furnham noted that Extraversion was most likely to correlate strongly with implicitness in formal situations, such as examinations. Weaker correlations were found in more informal (contextual) situations. Now, according to the results in the first part of this paper (see Table 2), blogs are more formal than e-mail. But they are still relatively informal—just surpassing School Essays in their F-score. Even ignoring the non-linearity just noted, this relative informality could reflect informality in the communication situation, and thus fit with the low correlation between Extraversion and F-measure found in the current corpus.

Heylighen and Dewaele also discussed the relation between Openness and F-measure, although at the time they had no corroborating evidence for a link. They hypothesised that since Openness is also considered the factor of intellect, it should correlate positively with formality. This is what we have found.

Agreeableness and language use, however, have not been extensively discussed previously. One aspect of Agreeableness is cooperativity: highly Agreeable individuals are most willing to cooperate and accommodate. In communication, this could be realised via a better ability—or at least willingness—to adapt to the interlocutor's communication situation or style. Interpreting this in the setting of blogs suggests that bloggers of an Agreeable nature are more likely to be aware of the lack of shared context between themselves and the reader, thus adjusting their writing away from contextuality, in the direction of formality.

Conclusion

The study has used Heylighen and Dewaele's F-measure to draw two main conclusions. One is that blogs are, as a genre, likely to prove more formal than e-mail, but not much more formal than School Essays. The other is that, within the blog genre, there is variability in contextuality/formality due to individual differences. But the differences that make the most difference are not Extraversion or Neuroticism, but Openness and—especially—Agreeableness and gender.

Acknowledgements

Our thanks to Judy Robertson for her helpful input. The first author gratefully acknowledges support from the UK Economic and Social Research Council, studentship number R42200134353.

References

Argamon, S., Koppel, M., Fine, J., & Shimoni, A. (2003). Gender, genre, and writing style in formal written texts. *Text, 23(3).*

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

Belo, R. (2004). *Blogs take on the mainstream.* BBC News Online. Available at http://news.bbc.co.uk/1/hi/technology/4086337.stm

Buchanan, T. (2001). *Online implementation of an IPIP Five Factor Personality Inventory [On-line].* Available at: http://users.wmin.ac.uk/∼buchant/wwwffi/introduction.html

Cohn, M.A., Mehl, M.R., & Pennebaker, J.W. (2004). Linguistic markers of psychological change surrounding September 11, 2001. *Psychological Science, 15,* 687-693.

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

Costa, P., & McCrae, R. R. (1992). *Professional Manual.* Psychological Assessment Resources, Odessa, FL.

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

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

Furnham, A. (1990). Language and personality. In H. Giles & W. Robinson (Eds.), *Handbook of Language and Social Psychology.* Wiley, Chichester.

Gill, A., & Oberlander, J. (2002). Taking care of the linguistic features of Extraversion. *Proceedings of the 24th Annual Conference of the Cognitive Science Society* (pp. 363–368). Hillsdale, NJ: LEA.

Gill, A., & Oberlander, J. (2003). Perception of e-mail personality at zero acquaintance: Extraversion takes care of itself; Neuroticism is a worry. *Proceedings of the 25th Annual Conference of the Cognitive Science Society* (pp. 456–461). Hillsdale, NJ: LEA.

Gill, A., Harrison, A. & Oberlander, J. (2004). Interpersonality: Individual differences and interpersonal priming. In *Proceedings of the 26th Annual Conference of the Cognitive Science Society,* (pp. 464–469). Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Heylighen, F., & Dewaele, J.-M. (2002). Variation in the contextuality of language: an empirical measure. *Foundations of Science, 7,* 293–340.

Keller, F., Lapata, M., & Ourioupina, O. (2002). Using the web to overcome data sparseness. In Jan Hajic and Yuji Matsumoto, (Eds.), *Proceedings of the Conference on Empirical Methods in Natural Language Processing,* (pp. 230–237). Philadelphia.

Lee, D. Y. W. (2001). Genres, registers, text types, domains and styles: clarifying the concepts and navigating a path through the BNC jungle. *Language Learning & Technology, Vol.5(3),* 37-72.

Levelt, W.J.M. (1989). *Speaking. From Intention to Article.* MIT Press, Cambridge, Mass..

Louwese, M., McCarthy, P.M., McNamara, D.S., & Graesser, A.C. (2004). Variation in language and cohesion across written and spoken registers. *Proceedings of the 26th Annual Conference of the Cognitive Science Society* (pp. 1035–1040). Hillsdale, NJ: LEA.

Markey, P. and Wells, S. (2002). Interpersonal perception in internet chat rooms. *Journal of Research in Personality, 36,* 134–146.

Matthews, G., Deary I., & Whiteman, C. (2003). *Personality traits: Second Edition.* Cambridge: Cambridge University Press.

Mortensen, T., & Walker J. (2002) Blogging thoughts: personal publication as an online research tool. In A. Morrison (Ed.) *Researching ICTs in Context,* InterMedia Report, 3/2002, Oslo.

Oberlander, J., & Gill, A. (2004). Individual difference and implicit language: personality, parts-of-speech and pervasiveness. *Proceedings of the 26th Annual Conference of the Cognitive Science Society* (pp. 1035–1040). Hillsdale, NJ: LEA.

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

Ratnaparkhi, A. (1996). A maximum entropy part-of-speech tagger. *Proceedings of the Conference on Empirical Methods in Natural Language Processing,* University of Pennsylvania.

Rainie, L. (2005). *The state of blogging.* Pew Internet & American Life Project. Available at http://www.pewinternet.org/PPF/r/144/report_display.asp

Rayson, P., & Leech, G., & Hodges, M. (1997). Social differentiation in the use of English vocabulary: some analyses of the conversational component of the British National Corpus. *International Journal of Corpus Linguistics, Vol.2(1),* 133-152.

Rosenbloom, M. (Ed.) (2004). The Blogosphere. *Communications of the ACM, Special Issue, 47 (12).* ACM Press, New York

Volk, M. (2001). Exploiting the WWW as a corpus to resolve PP attachment ambiguities. In P. Rayson, A. Wilson, T. McEnery, A. Hardie, & S. Khoja, (Eds.), *Proceedings of the Corpus Linguistics Conference,* 601-606, Lancaster.