The Identity of Bloggers: Openness and gender in personal weblogs

S. Nowson and **J. Oberlander**
School of Informatics, University of Edinburgh
2 Buccleuch Place, Edinburgh, EH8 9LW
s.nowson,j.oberlander@ed.ac.uk

Abstract

Work has recently been completed on a PhD Thesis concerning individual difference in the language of personal weblogs (Nowson 2005). This paper highlights some of the results. Blogs are increasingly used as a resource for academic study, as evidenced by this symposium. Bloggers are not, however, representative of the population as a whole: they are more likely to be teenage or 20-something females, and appear to be highly Open to Experience. Following our linguistic analysis of personal blog entries, we are constructing a feature set for the automatic detection of gender: a small amount of n-gram context proves best at accounting for variance when compared to dictionary-based analysis.

Introduction

In the context of this symposium weblogs need little introduction. This paper is specifically concerned with personal diary-style weblogs, or 'blogs' as they are more commonly known. The paper designed to give a flavour of the research we have carried out: a linguistic analysis of gender and personality differences in the language of blogs (Nowson 2005).

A corpus of personal weblogs was constructed. Participants were recruited via direct mailing of suitable candidates and by word-of-mouth: many participants wrote about the study in their blogs. The first task of the study was to complete sociobiographic and personality questionnaires. The inventory chosen was specifically created and validated so as to be completed online (Buchanan 2001). It was derived from the 50-item IPIP implementation of five factors of Costa & McCrae's (1992) revised NEO personality inventory. It consists of 41-items on which participants rate themselves using a 5-point Likert scale.

The text corpus was created by the second task of the study: that each participant submit one month's worth of prior blog postings. The month was pre-specified so as to reduce the effects of an individual choosing their 'best' or 'favourite' month. Raw submissions were marked-up using XML in order to automate extraction of the desired text. The tagset included high level features such as *posttitle* and *date* along with low level features such as *link* and *image*. Text was also marked corresponding to which type of blog post it was: for example purely personal, commentary on external matters, or direct posting of internet memes such as quizzes. The corpus consisted of 71 participant's and when the text considered 'personal' was extracted, it approximated 410,000 words.

Amongst the work we have completed is a replication of the Linguistic Enquiry and Word Count (LIWC; Pennebaker & Francis 1999) factor analysis work of Pennebaker and King (1999). We have shown that the underlying factor structure for blogs is similar to the structure the original study found in personal student essays. We have also used a measure of contextuality (Heylighen & Dewaele 2002) to place blogs in the context of other genres (also reported in Nowson, Oberlander, & Gill 2005). We found blogs to be: less contextual than e-mails and personal letters; similar to high school level essays; and more contextual than journalistic and academic writing.

This paper introduces two aspects of the larger work. Firstly we discuss our findings as to bias within the demographics of the blogging population. This should be of interest to any researchers wishing to construct a balanced blog corpus. Secondly, we employed a number of analytical approaches in order to investigate linguistic differences due to both personality and gender. Here, we will highlight some of the findings for gender, and detail how these are being used to construct a feature set for automated gender detection.

Blog Demographics

Though it was not an aim of the original work (Nowson 2005), examination of the sociobiographic data collected as part of the blog corpus revealed a number of interesting biases. These concerned both personality traits, gender and, to a lesser extent, age. This finding should be of interest to researchers. While this corpus was hand collected, it is tempting to automatically create large corpora by harvesting blog lists. Without profile information, however, it is difficult to know the demographic makeup of such a population. As such it is hard to create a balanced corpus containing no bias towards groups within it.

Gender and age

Diaries are traditionally more likely to be kept by females (Thompson 1982; Burt 1994) and it has previously been suggested that the majority of bloggers are teenage

girls (Orlowski 2003). In fact a recent study (Herring *et al*. 2004) found that while each gender accounts for about half of all weblogs, *personal* blogs are dominated by females of teen age and preferred by females in general.

Indeed, of the 71 bloggers in our corpus, there were 47 females and 24 males, with average ages 27.8 and 29.4 respectively. These are similar averages but the distribution of participants within age ranges differs between genders. Figure 1 shows the number of both male and female participants, along with the total, that lie within each age range.

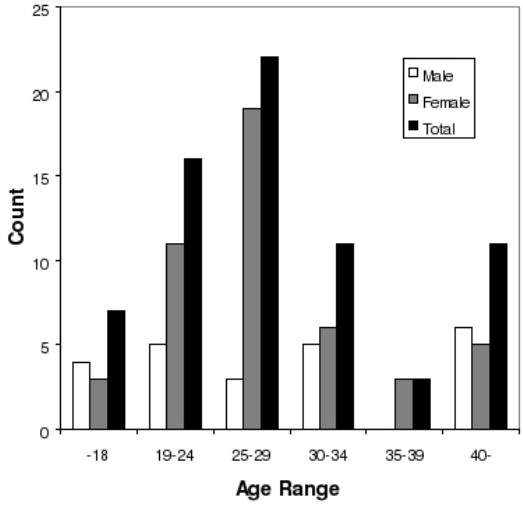

Figure 1: The number of participants within each age range, by gender and in total

Whilst teenagers are not the majority group here, participants under 30 clearly dominate. These younger groups also consist of many more females than they do males. This would seem to support earlier suppositions that there are more young female bloggers than any other sub-group.

However, it is not just the raw number of men and women bloggers that could have a biasing effect on a corpus. Previous studies of CMC (Herring 2000) have found that women generated shorter posts than men to discussion lists and news groups. However, these studies have tended to focus on work-based groups, and findings also suggest that women are less confident posting, often feeling intimidated by their male colleagues. Blogs, by contrast, are personal and individually written, closer to situations in which women were reported to participate to a greater extent.

Our blog corpus consists of approximately 410,000 words. The 47 females and 24 males provide approximately 320,000 and 90,000 words across 1400 and 450 posts respectively. This means that while the men wrote an average of about 3700 words in the month, females wrote 6800 each, almost double (despite large standard deviations, this is a significant difference: $t(69) = 2.315$, $p < .05$). This is reflected not only in longer posts, but in more frequent postings: women wrote on average 30 personal posts in a month (SD 22.7) while men wrote only 19 (SD 11.6; this is also a significant difference: $t(69) = 2.196$, $p < .05$); women wrote an average of 251 words in each post (SD 186), while men wrote 194 words (SD 117; a non-significant difference).

It is clear therefore, that not only are personal weblogs dominated by younger females, but that females also spend more effort blogging than men. This is interesting for academics wishing to automatically cull a blog corpus from the internet. It is likely that the corpus will be dominated by females, not only in number, but more importantly in word count. Gender cannot always be determined automatically from blogs. Herring et al. (2004) determined the gender of 94% of the authors in their collection by using:

> names, graphical representations (if present), and the content of blog entries (e.g., references to "my husband" resulted in a "female" gender classification, assuming other indicators were consistent).

In our study author information was carefully collected, requiring questionnaires from participants in which they specified their gender.

Personality

It is not just in gender and age that there is a bias: our studies have also investigated personality. The most intuitive bias one would expect in bloggers relates to their level of Extraversion. There is much written on the web that discusses the idea of keeping a blog as 'exhibitionism' and 'mental masturbation.'[1] This suggests that bloggers are perceived as Extraverts. This is plausible, since bloggers describe details of their lives to anybody who is interested. This is consistent with Extraverts' greater desire to communicate (McCroskey & Richmond 1990).

A plausible counter argument, however, concerns the perceived anonymity of the online world. Bloggers write at a distance from their readers and can conceal their identity if they so wish. This suggests that bloggers may be Introverted by nature. This too seems plausible since it is easy to imagine Extraverts confiding with friends directly about their activities and thoughts, while Introverts might prefer to communicate through a potentially anonymous written medium. In addition to this, a recent study has found that one common motivation for blogging is the author's desire to be a writer (Li 2005): writers are generally perceived to be Introverts, due to the solitary nature of the working environment.

By plotting the distribution of Extraversion scores we can test this hypothesis. Our study concluded that in fact bloggers are neither Extrovert nor Introvert. It is clear from figure 2 that whilst there does appear to be a slight Extravert bias, the distribution is a good approximation of the normal curve, with a wide range of scores.

The same can also be said of Neuroticism, Agreeableness and Conscientiousness, though there appears to be a slight bias in favour of high Agreeableness. Openness, however, is another matter. The two main observations that can be made from figure 3 are that the data is both unevenly distributed and significantly not normal. The lowest possible score on the Openness scale is 7, yet the lowest score from the blog

[1] See `http://jilltxt.net/?p=85` for discussion and rebuttal.

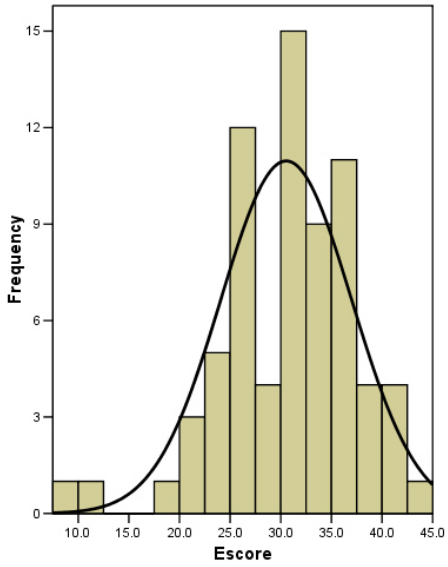

Figure 2: Distribution of Extraversion scores

data is 18, which only one participant scored. Conversely, 10 participants scored the maximum 35 points, with 8 scoring 34.

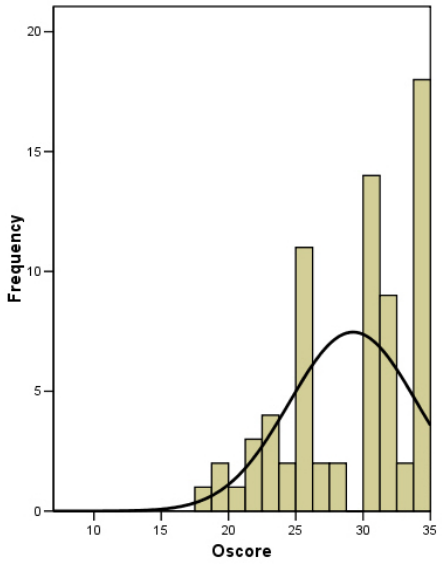

Figure 3: Distribution of Openness scores

Ignoring the unevenness, it is clear there is a very heavy bias toward scoring high on Openness. This may be indicative of the nature of bloggers, that they are very Open individuals. It certainly seems plausible that people who post details of their lives online could be described as open. Of course, it is possible that this result only reflects the bias among those who chose to submit data. Participation required participants to submit personal details and it is easy to imagine that only the most open of individuals are prepared to do so.

From personal correspondence with Tom Buchanan, one of the authors of inventory used here, it is clear that the 'Open blogger' hypothesis will remain just that. Whilst it is plausible, without a comparison sample such as non-bloggers recruited and tested in the same way, the exact cause of this anomaly cannot be determined.

Towards Gender Identification

The bulk of our work has employed linguistic analysis to identify language differences due to individual differences. Whilst this has also been carried out for the personality traits, here we discuss just the gender findings. Investigation has revealed a number of features that appear to lend themselves well to use for distinguishing between the language of males and females.

There is much previous work on gender identification through language, both automatic (Koppel, Argamon, & Shimoni 2002; Argamon *et al.* 2003) and otherwise (see Herring 2000, Pennebaker, Mehl, & Neiderhoffer 2002 for reviews). Gender language differences have also been previously investigated in blogs (Huffaker 2004).

The first approaches we employed were dictionary-based content analyses. We used both the LIWC, which has previously been used to investigate language in blogs (Cohn, Mehl, & Pennbaker 2004) and following Gill (2004), a novel application of the MRC Psycholinguistic database (Coltheart 1981; Wilson 1987). Examining those variables which showed at least a marginally significant relationship with gender ($p < .1$) confirmed prior findings: women have a more immediate style of writing, use more pronouns and words reflecting social processes, emotions and physical states; men by contrast talk more about external matters, they use more articles and (according the MRC) more frequent, less concrete language.

We then turned to more data-driven approaches. The first was the measure of contextuality mentioned earlier (Heylighen & Dewaele 2002). This was a measure derived from the notion of deixis: that some parts-of-speech (POS) generally (though not invariably) require more context in order to disambiguate them than others. For example, the pronoun in '*get it*' is easier to resolve with context such as '*see the ball by the front door, get it*'. Women were shown to have a significantly more contextual style of writing than men ($t(66) = -2.75$, $p < .01$). This difference also holds in other genres such as fiction and email, though not in the rather formal field of academic writing, when the genders score similarly (Nowson, Oberlander, & Gill 2005). Since this measure is based POS, we could also see that once again men use more articles and fewer pronouns, but also more nouns, prepositions and fewer verbs.

We also examined context as reflected in the use of distinctive word collocations. By using the G^2 log-likelihood statistic (Dunning 1993), bi- and trigrams were identified which occurred with significant relative frequency distinctly in the language of one gender over the other. Male n-grams appear to consist of mostly function word combinations (e.g. [*on a*] and [*of the*]), while females were characterised by

Feature Set	No.	Accuracy	Precis.	Recall
LIWC	19	77.5%	.749	.749
MRC	7	63.4%	.326	.479
N-grams	125	93.0%	.915	.923
All	157	**91.5%**	**.901**	**.916**

Table 1: SVM classification of gender

phrases with content words. There were also many more n-grams containing pronouns. The n-grams were selected on the basis that they occurred in at least 1/3 of the participants in each group (8 males and 15 females). Individual influence was then further reduced by determining those n-grams that showed a statistically significant relationship with gender.

In order to compare the effectiveness of the features our studies had revealed, we have conducted simple machine classification experiments. A support vector machine (SVM) approach was used to classify author gender. SVMs have previously been shown to be generally better than Naïve Bayes (Joachims 1998). As described above, the features chosen were those variables which had previously shown at least a marginally significant relationship with gender.

Similar features such as n-grams and POSs have been used previously in automatic text classification. Pang, Lee, & Vaithyanathan (2002) compared approaches, including SVM, to classify film review sentiment as positive or negative. They also compared bag-of-features sets including unigrams, bigrams and POS tags, which varied in size between 3000 and 32000 features. Argamon and Koppel (Koppel, Argamon, & Shimoni 2002; Argamon *et al.* 2003) have employed both lists of function words and POS n-grams to begin with over 1000 potential features in the classification of gender. Their approach however enabled them to select those features most useful for classification. Reducing their features set initially improved results. However, there was only a 5% loss of performance overall with as few as 8 features. As already discussed, it is our previous work that informs our feature selection.

The WEKA work bench was used to implement 10-fold cross validation studies on the 71 datafiles. Accuracy results, from the features sets of this work, alongside average precision and recall scores, can be seen in table 1.

Note that due to the imbalance in gender in the corpus the base rate for this study is 66.2%, achieved simply by classifying all participants as females. The lower than base rate score from the MRC features is due to two females being the only participants classfied as male. It is also worth noting that the experiment reported here is simply an initial starting point for our work in this direction; there has been no attempt to refine the classifier or feature set. There is also room for overfitting since the feature set is used to classify the very data from which it is derived.

However, despite these obvious limitations, it is clear that using the context-based n-gram set appears more accurate than using either of the dictionary-based approaches. To control for the simple fact that there are a greater number of n-gram features, the set was reduced by removing those feature with lowest significance. The results of the n-gram only features sets can be seen in figure 4. As with previous studies, small feature sets can still prove remarkably accurate for classification.

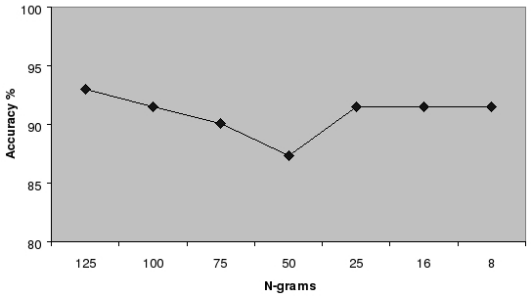

Figure 4: Accuracy of reduced n-gram sets

Discussion

The dominance of context-based n-grams is echoed in our studies and results for personality. It is planned that these studies will provide the basis for feature sets for automatic classification of not just gender but personality classes. We intend to further investigate differing approaches to machine classification, as well as alternative feature selection methods to our own empirical methodology. It is also hoped that our so far promising results will be replicated with the use of unseen data.

Final words

This paper is intended to give just a flavour of our work on the language of personal weblogs. It has highlighted apparent biases in the demographics of the weblog community, which are worth noting when constructing a corpus. It has also briefly introduced our work leading towards automatic classification of blogger gender from text.

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.; and Shimoni, A. 2003. Gender, genre, and writing style in formal written texts. *Text* 23(3).

Buchanan, T. 2001. Online implementation of an IPIP five factor personality inventory [on-line]. Retrieved 25th October, 2005, from http://users.wmin.ac.uk/~buchant/wwwffi/introduction.html.

Burt, C. 1994. Prospective and retrospective account-making in diary entries: A model of anxiety reduction and

avoidance. *Anxiety, Stress & Coping: An International Journal* 6(4):327–340.

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

Coltheart, M. 1981. The MRC psycholinguistic database. *Quarterly Jour. of Experimental Psychology* 33A:497–505.

Costa, P., and McCrae, R. 1992. *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI): Professional Manual*. Odessa, FL: Psychological Assessment Resources.

Dunning, T. 1993. Accurate methods for the statistics of surprise and coincedence. *Computational Linguistics* 19(1):61–74.

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

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

Herring, S.; Kouper, I.; Scheidt, L.; and Wright, E. 2004. Women and children last: The discursive construction of weblogs. In Gurak, L.; Antonijevic, S.; Johnson, L.; Ratliff, C.; and Reyman, J., eds., *Into the Blogosphere; Rhetoric, Community and Culture of Weblogs*. University of Minnesota. Retrieved October 25th, 2005, from http://blog.lib.umn.edu/blogosphere/women_and_children.html.

Herring, S. 2000. Gender difference in CMC: findings and implications. *Computer Professionals for Social Responsibility Journal* 18(1).

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

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.

Joachims, T. 1998. Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of the European Conference on Machine Learning (ECML)*, 137–142.

Koppel, M.; Argamon, S.; and Shimoni, A. 2002. Automatically categorizing written texts by author gender. *Literary and Linguistic Computing* 17(4):401–412.

Li, D. 2005. Why do you blog: A uses-and-gratifications inquiry into bloggers' motivations. Master's thesis, Marquette University.

McCroskey, J., and Richmond, V. 1990. Willingness to communicate: A cognitive view. *Journal of Social Behaviour and Personality* 5:19–37.

Nowson, S.; Oberlander, J.; and Gill, A. 2005. Weblogs, genres and individual differences. In *Proceedings of the 27th Annual Conference of the Cognitive Science Society*, 1666–1671. Hillsdale, NJ: Lawrence Erlbaum Associates.

Nowson, S. 2005. *The Language of Weblogs: A study of genre and individual differences*. Ph.D. Dissertation, University of Edinburgh.

Orlowski, A. 2003. Most bloggers "are teenage girls" - a survey. The Register. Retrieved October 25th, 2005, from http://www.theregister.co.uk/content/6/30954.html.

Pang, B.; Lee, L.; and Vaithyanathan, S. 2002. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 79–86.

Pennebaker, J., and Francis, M. 1999. *Linguistic Inquiry and Word Count: LIWC*. Erlbaum Publishers.

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

Pennebaker, J.; Mehl, M.; and Neiderhoffer, K. 2002. Psychological aspects of natural language use: Our words, our selves. In *Annual Review of Psychology*, volume 54. 547–577.

Thompson, E. 1982. *The Makings of the English Working Class*. Harmondsworth: Penguin.

Wilson, M. 1987. MRC psycholinguistic database: Machine usable dictionary. Technical report, Oxford Text Archive, Oxford.